

Mental Testing

ITS
HISTORY,
PRINCIPLES,
AND
APPLICATIONS

by

FLORENCE L. GOODENOUGH

Institute of Child Welfare • University of Minnesota

RINEHART & COMPANY, INC., *Publishers • New York*

First Printing September, 1949
Second Printing June, 1950
Third Printing February, 1953
Fourth Printing April, 1954

Copyright, 1949, by Florence L. Goodenough

Printed in the United States of America

All Rights Reserved

Designed by Stefan Salter

TO
LEWIS M. TERMAN
A WORTHY SUCCESSOR TO ALFRED BINET

Preface

More than half a century has passed since the appearance in 1890 of Cattell's classic article in which the term "mental test" was introduced into the psychological literature. Could the author then have foreseen the extraordinary growth and development which his idea was destined to attain within his own lifetime it is probable that no one would have been more amazed than he. Few if any other fields of psychology have aroused so widespread an interest, not only among psychologists but within such related fields as psychiatry, anthropology, sociology, and education and in certain branches of the law, as well as among the general public. It was perhaps inevitable that so rapid and widespread a growth of interest has led to the promulgation of erroneous as well as sound theories, and that the use of the new methods by many enthusiastic but poorly trained persons has not always worked out to the advantage of the tested individuals or of society. Young sciences, like young learners, must try out many pathways in order to discover the true course, and in both cases the early stages of learning are likely to be marked by overhasty conclusions and wishful thinking.

But mental testing may now be fairly said to have attained its majority with the prospect of a long and active maturity lying ahead. It seems advisable, then, to take time to survey the possibilities and limitations of the method and to consider the theories and assumptions, both explicit and implicit, underlying the construction, administration, and interpretation of tests. In the hope of avoiding like mistakes in the future, some of the erroneous conclusions drawn by our predecessors will be pointed out, while for the encouragement of young adventurers in the field a number of promising avenues that hitherto have been overlooked or incompletely explored will be mentioned. In order that the present status of the testing movement may be more clearly understood, the first few chapters will present a brief historical overview both of the social problems that showed the need for tests and of the progress in scientific knowledge and thought that gave rise to the idea and provided for its implementation.

This book is planned to serve the needs of several groups. Many years of experience, both in training students for psychometric and clinical practice and in working with schools, social agencies, and other organizations which endeavor to make use of tests in the practical

handling of individuals, have convinced me that a basic defect in the scientific background of a large number of the persons at present engaged in testing is their lack of understanding of the theoretical principles underlying the procedures which they employ. I have accordingly tried to indicate the nature of these principles as they apply to the actual testing of individuals, and in doing so have employed, as far as possible, simple and nontechnical language. Such technical terms as it has been impossible to avoid are explained in the Glossary (pages 543-569). The book is thus designed to serve as an orientation text for students planning to enter the field of testing. Clinical psychologists and psychometrists may also find that it contains material of interest to them, since the organization and treatment differs considerably from that of most previous books on this subject.

The second major group of readers for whom the book is designed is composed of the large body of professional workers who make use of the results of tests which they do not themselves administer. School principals, teachers, social workers, psychiatrists, and pediatricians, juvenile court judges, and many others are daily faced with problems in which the question of the intellectual and social competence of children or adults is involved. The greater number of these persons, recognizing the potential value of test results as an aid to the effective handling of their clientele, are nevertheless hampered in their utilization of such material because of their lack of understanding of the procedures used and their ignorance of the many pitfalls into which the well-meaning but uninformed practical worker who tries to handle test results by rule of thumb is all too likely to stumble. This book has been written in the hope that the discussion which follows will help those who are concerned with the training and guidance of human beings to a better understanding of the appraisal of mental abilities as an aid to social welfare.

In the use of the book some previous training in statistical methods is desirable, but not essential. The measurement of ability and the appraisal of behavior alike demand a quantitative approach, and the success of each depends upon the appropriateness of the mathematical procedures by which the crude and fluctuating results of casual observation and personal opinion are transformed into the relatively precise and objective standards of the mental test. For those who are actively engaged in the derivation of new tests or in the critical examination and evaluation of those already available, a thorough acquaintance with statistical method and the mathematical principles upon which it is based is essential. For the many practical workers who attempt to use tests as an aid to the understanding and guidance of human beings, some

knowledge of statistical terminology, of the meaning of common statistical formulas and the conditions required for their use is also needed if misinterpretations are to be avoided and the potential value of the tests is to be completely realized. It is primarily for this group that the methodological discussions in Part II have been written. It should be emphasized that this section is by no means to be looked upon as the equivalent of a short course in statistical method. Its only purpose is to enable the reader without technical training to make more intelligent use of test results through better understanding of the basic assumptions upon which tests have been developed and of their consequent possibilities and limitations.

In conclusion I should like to express my gratitude to the many authors and publishers who have granted permission for the reproduction of material from their works. Special thanks are due to the Educational Test Bureau of Minneapolis for allowing me to reprint in full the suggestions on the "Conduct of an Examination" from the *Manual of instructions for the Minnesota Preschool Scales*; to Professor Ronald A. Fisher and to Messrs. Oliver & Boyd Ltd., Edinburgh, for permission to reprint Tables III and IV from their book, "Statistical Methods for Research Workers"; to Houghton Mifflin Company for permission to quote from Lewis M. Terman's *The intelligence of school children* and from Howard C. Warren's *Dictionary of psychology*; and to Professors G. W. Snedecor and E. F. Lindquist for permission to reproduce portions of their statistical tables.

F.L.G.

Lisbon, New Hampshire
July 1949

Contents

PREFACE	PAGE vii
LIST OF ILLUSTRATIONS	xvii
LIST OF TABLES	xix

PART I · HISTORICAL ORIENTATION

CHAPTER 1 · THE SOCIAL NEED FOR MENTAL DIAGNOSIS	3
Identification and classification of the feeble-minded—American interest in mental defect—The young offender	
CHAPTER 2 · THE EDUCATIONAL NEED	14
CHAPTER 3 · THE SCIENTIFIC BACKGROUND	20
Early attempts at describing intelligence—The rise of abnormal psychology—Mental inheritance in man—Statistical methods—Advances in medical knowledge—The psychology of sensation and its relation to mental testing—The child study movement	
CHAPTER 4 · THE EARLY TESTS. 1887-1915	34
Informal attempts at appraising intelligence—The search for bodily signs—Early experiments with tests of a single kind: Sensorimotor tests—Single tests of a more complex nature—The 1905 scale—The 1908 scale—The 1911 scale—The Use of the Binet-Simon Tests in America: Goddard's translations—Mental tests and heredity	
CHAPTER 5 · LATER DEVELOPMENTS	59
Early attempts at revising the Binet-Simon scale—The Stanford 1916 Revision—The Stanford 1937 Revision—World War I and the advent of group testing—The development of nonverbal tests—Educational tests—Tests of motor and perceptual abilities—Tests for industrial selection and vocational guidance—The appraisal of "personality"—"Projective methods for assaying the personality"—Advances in statistical method	

	PAGE
CHAPTER 6 • THE PRESENT STATUS	89
Rapid growth of the testing movement—Some common fallacies regarding tests and testing	
<i>PART II • PRINCIPLES AND METHODS</i>	
CHAPTER 7 • THE BEARING OF TESTING THEORY UPON TEST INTERPRETATION	97
Defining a universe from signs or samples—The prediction of behavior from the study of signs—The prediction of behavior from the study of samples—The question of trait names—The experiential reference	
CHAPTER 8 • PROBLEMS OF SAMPLING	109
The test as a sample of abilities or conduct—Methods of sampling: I. Random sampling—Methods of sampling: II. Stratified sampling—The standard sample as a source of reference—Defining the universe to be sampled—The question of sampling in test interpretation	
CHAPTER 9 • THE ANALYSIS AND SELECTION OF TEST ITEMS	123
Tridimensional aspect of most behavior traits—The selection of test items—The analysis and selection of items in tests of the limited-response type—Some practical applications	
CHAPTER 10 • UNITS OF MEASUREMENT	140
Reference standards and interpretative measures—Item counts and “absolute scaling”—The question of the zero point—The effect of insufficient range of difficulty upon test scores—Chronological age as a basis for scaling—Mental growth curves—Psychological measures compared with physical measures	
CHAPTER 11 • AGE STANDARDS AND QUOTIENTS AS INTERPRETATIVE MEASURES	157
The need for uniformity of quantitative terminology—Mental age as an interpretative measure—The intelligence quotient—Factors affecting the meaning of an intelligence quotient—Some reasons for IQ constancy	
CHAPTER 12 • MEANS, MEDIANS, AND PERCENTILES	177
The choice of an average in group comparisons—Percentile ranks as interpretative measures—The stability of percentiles—Other measures based on percentages	

	PAGE
CHAPTER 13 • STANDARD SCORES AND THEIR DERIVATIVES	189
Meaning and derivation of standard scores—Other ways of expressing standard scores: The T-Score method—The “discriminative value” method of Arthur and Woodrow	
CHAPTER 14 • OTHER DEVICES FOR INTERPRETING TEST SCORES	204
The Heinis Personal Constant (PC)—The per cent placement—The median mental age—Other interpretative measures—Comparative note on interpretative measures	
CHAPTER 15 • TESTING THE TESTS. I. GENERAL PRINCIPLES AND FUNDAMENTAL METHODS	213
Checks to be applied—The criterion measure—The relative importance of the stability of test results and of the validity of the conventional interpretations of test meaning—Multiple criteria—Factor analysis and the “purification” of tests—Concluding remarks	
CHAPTER 16 • TESTING THE TESTS. II. THE DIVERGENCE OF FACTS FROM HYPOTHESES	232
The null hypothesis—Degrees of freedom—Chi-square (χ^2) and the null hypothesis—Differences between means of continuous variates—The practical use of tests of “significance” in mental measurement	
CHAPTER 17 • CORRELATED MEASURES	251
Bases of correlation—Dependent and independent variables—Some correlates of intelligence—Further methods of determining relationships	
CHAPTER 18 • ANALYSIS OF VARIANCE	271
Basic concepts—Practical applications—The design of experiments	
CHAPTER 19 • SOME QUESTIONS OF MENTAL ORGANIZATION	286
Explicit and implicit hypotheses—The mathematical analysis of nonintellectual traits	

PART III • TESTS AND SCALES

CHAPTER 20 • THE CONDUCT OF AN EXAMINATION, WITH PARTICULAR REFERENCE TO THE TESTING OF YOUNG OR DIFFICULT CHILDREN	297
General notes—The conduct of an individual examination of a young child—Individual examinations of older children and adults—Notes on group testing	

	PAGE
CHAPTER 21 • INTELLIGENCE TESTS	308
Tests for infants—Tests for children of preschool age—Tests for kindergarten children—Tests for elementary school children—Tests for high school and college students—Tests for unselected adults—Mental tests in senescence and old age	
CHAPTER 22 • TESTS OF EDUCATIONAL APTITUDE AND ACHIEVEMENT	322
“Readiness” tests—Tests of educational achievement in the elementary school—Educational tests for use in secondary schools and colleges—The comparison of ability with accomplishment—Practical applications of educational measurement	
CHAPTER 23 • THE MEASUREMENT OF SPECIAL TALENTS AND DEFICIENCIES	338
The relation of special talent to general intellectual ability—Methods of study—Measures of special talent—Craftsmanship versus creativeness—“Idiots savants”—Specialized mental or educational deficiencies—Other diagnostic measures—Some practical considerations	
CHAPTER 24 • THE MEASUREMENT OF MOTOR DEVELOPMENT AND MOTOR SKILL	360
The concept of general motor ability—The measurement of motor development in infancy and early childhood—Motor development in later childhood and adolescence—Motor tests for college students and for unselected adults—The measurement of muscular strength—Tests of lateral dominance—The correlation of motor skill with other abilities	
CHAPTER 25 • THE MEASUREMENT OF INTERESTS AND ATTITUDES	374
Purposes and aims—Methods of studying interests—Practical applications of the measurement of interests—Methods of studying attitudes—Practical applications of the measurement of attitudes—Sources of error in the measurement of interests and attitudes—The public opinion poll	
CHAPTER 26 • THE MEASUREMENT OF PERSONAL—SOCIAL CHARACTERISTICS	390
Definitions—General methodological considerations—Methods of studying social participation—Methods of studying social organization—Other methods of studying social relationships—Standardized rating scales for appraising social behavior—Paper-and-pencil	

tests and questionnaires for the appraisal of personal-social characteristics

CHAPTER 27 • PROJECTIVE METHODS FOR THE STUDY OF PERSONALITY 415

Fundamental concepts—Binet's *Experimental Study of Intelligence*—Doll play—The interpretation of art products—Other methods used principally with children—Methods used chiefly with adolescents and adults. Free word association—Murray's Thematic Apperception Test—The Rorschach Inkblot Test—Other diagnostic methods based upon the theory of projection—Critical discussion of the projective theory and methodology

CHAPTER 28 • TESTS FOR VOCATIONAL GUIDANCE 442

The aims of vocational guidance—The place of vocational counseling—Qualifications of the vocational counselor—Tests for vocational guidance—Tests of special vocational abilities and aptitudes

PART IV • APPLICATIONS

CHAPTER 29 • TESTING IN SCHOOLS AND COLLEGES 455

The major objective—The applications of measurement to educational purposes

CHAPTER 30 • TESTING IN CLINICAL PRACTICE 459

Variable factors affecting the selection and use of tests—Psychological testing in psychiatric clinics for adults—Psychological testing in behavior clinics for children—The role of the psychologist in rehabilitation work with adults and older adolescents—Mental testing of physically handicapped children—Special problems of the clinical psychologist

CHAPTER 31 • THE USE OF TESTS IN INDUSTRY 471

Purposes for which tests are used—Tests for industrial selection and classification—The validation of tests for industrial purposes—The qualifications of the industrial psychologist

CHAPTER 32 • TESTING AND SOCIAL WELFARE 483

Psychology and medicine—The use of mental tests by social welfare agencies—The problem of the unmarried mother—The prevention and treatment of juvenile delinquency

	PAGE
CHAPTER 33 • TESTING THE ARMED FORCES	494
The use of tests in World War I—Psychological examining in World War II—The use of tests in the Army Air Forces—Psychological testing in other branches of the service—A comparison of military psychology in 1941–1945 with that of 1917–1918	
CHAPTER 34 • TESTING AND SCIENTIFIC INVESTIGATION	504
The mental test as an interdisciplinary tool of research	
CHAPTER 35 • THE USE OF TESTS IN THE STUDY OF GROUP DIFFERENCES	510
The scientific value of comparative studies—Extreme deviates in intellectual ability—Delinquents and criminals—Sex differences—National-racial differences—Mental differences among cultural groups—The physically handicapped	
CHAPTER 36 • LOOKING AHEAD	532
The road behind us—Some unanswered questions—What of the future?	
GLOSSARY	543
BIBLIOGRAPHY	571
AUTHOR INDEX	593
SUBJECT INDEX	597

List of Illustrations

	PAGE
Frontispiece—Portrait of Lewis M. Terman	
FIGURE	
1. Francis Galton, aged 87, on the stoep at Fox Holm, Cobham, with his biographer, Karl Pearson	26
2. Portrait of Alfred Binet	47
3. Some American pioneers in mental testing—Cattell, Yerkes, Goddard, Thorndike	60
4. The normal probability curve	147
5. Skewed curve resulting from the use of a test that is too difficult for the group to whom it is given	148
6. Skewed curve resulting from the use of a test that is too easy for the group to whom it is given	148
7. Four main types of postnatal growth	159
8. A percentile curve	183
9. Unequal distances between points subtended on the base line of a normal curve by successive 10 per cent divisions (deciles) of its area	186
10. Equal distances between points subtended on the base line of a rectangle by successive 10 per cent divisions (deciles) of its area	186
11. Showing the date of Table 3 after the values have been transmuted into the form of a normal distribution	192
12. The curve of mental growth as calculated by Heinis from data reported by Vermeylen	205
13. Nonoverlapping scores of two relatively homogeneous groups	240
14. Overlapping scores of two more heterogeneous groups whose mean scores differ by the same amount as that shown in Figure 13	241
15. Correlation between scores on two forms of an interest test when $r = +.60$	258
16. Distribution of scores on a scatter diagram when $r = +1.00$	259
17. Curvilinear regression shown by scores on a test which is too difficult for the majority of children before the age of ten or eleven years	260
18. Some of the statisticians who have made notable contributions to the field of tests and measurements—Kelly, Thurstone, Guilford, Fisher	284

	PAGE
19. Schematic representation of the views of Spearman and Thorndike on the organization of intelligence	289
20. Some of the authors of modern intelligence tests—Gesell, Ball, Arthur, Wechsler, Merrill	309
21. Some of the materials used in the Minnesota Preschool Scales	313
22. Two of the items from the test of Primary Mental Abilities for Ages Five and Six by Thelma Gwynn Thurstone and L. L. Thurstone	314
23. Some of the tests comprised in the Arthur Point Performance Scale	317
24. Psychogram showing the standing of a poor reader on tests of intelligence and of school achievement	341
25. "Youth Imploring." A clay statue modeled by a congenitally blind youth of seventeen years	347
26. Early motor skills	361
27. Development of manual prehension in infancy	363
28. Motor performances in adolescence	366
29. A hand dynamometer	368
30. Specifications for constructing a manoptoscope for determining eye dominance	372
31. Illustrating the manner of using the manoptoscope described in Figure 30	372
32. Differences in the number of sympathetic social contacts made and received by two nursery school children: diagram of individual roles in the group	395
33. Sociogram showing choices of roommate by each of twenty boys in a small private school	400
34. Binet's two daughters	421
35. Diagnostic signs in the motor response to words having emotional significance for the subject	429
36. An inkblot, similar to but not identical with those used in the Rorschach test	434
37. Two items from the Minnesota Paper Form Board Test	476
38. Two items from the Detroit Mechanical Aptitudes Examination	476
39. Two items from the O'Rourke Mechanical Aptitude Test	477
40. Correct and incorrect representation of the distribution of the intellectually inadequate in relation to mental test score	520
41. Bright children maintain their interest in a monotonous task by varying their manner of performing it	522
42. Drawings by feeble-minded children	524

List of Tables

TABLE	PAGE
1. Age-grade distribution of 11,769 children in the elementary schools of Dayton, Ohio, during the school year 1912-1913	16
2. Number and percentage of children retarded in school by three or more years in certain American cities in December, 1908, as reported by the U.S. Bureau of Education	17
3. Distribution of Canadian incomes for the year 1942	192
4. Table of the normal probability integral at successive 1 per cent levels	195
5. Comparison of IQ and PC values at three age levels	208
6. Chi-square (χ^2) values for different degrees of freedom	237
7. Levels of confidence for values of t at successive degrees of freedom	244
8. Values of F at the 1 per cent and the 5 per cent levels of confidence for differing degrees of freedom	248
9. Values of Fisher's z function for successive values of r	268
10. Distributions of scores on a test of mechanical ability by students from different high schools	275
11. Analysis of variance from data of Table 10	275
12. Distribution of scores made by women of differing marital status on a test of mental masculinity-femininity	280
13. Analysis of variance	280

Mental Testing

-

.

-

PART I

Historical Orientation

The Social Need for Mental Diagnosis

IDENTIFICATION AND CLASSIFICATION OF THE FEEBLEMINDED

Recognition of the social problems arising from the presence in society of those mentally incapable of meeting its needs has presumably existed to some extent from very early times, but there is little mention of it in the scientific literature until about the beginning of the nineteenth century. Previous to this time there was no very clear distinction between mental deficiency and mental disease except in cases of mania. Not until the appearance in 1838 of Esquirol's classic book on mental disorders was the basis for differentiating between the two made really explicit. Said Esquirol:

Idiocy¹ is not a disease but a condition in which the intellectual faculties are never manifested or have never been developed sufficiently to enable the idiot to acquire such amount of knowledge as persons of his own age reared in similar circumstances are capable of receiving. Idiocy begins with life or at that age which precedes the development of the intellectual and affective faculties, which are from the first what they are doomed to be during the whole period of their existence. Up to the present time no way has been found of altering this condition.

The mentally deranged person has been bereft of the good things which he once enjoyed; he is the poor man who once was rich; the idiot has always lived in misfortune and poverty. A state of abnormality may possibly be changed but the idiot remains always the same. The one [the idiot] has many childish features; the other [the mentally deranged] retains for the most part the physiognomy of an adult. Although both have little or no understanding, the abnormal show in their organization and even in their intelligence something of what they had previously attained; the idiot, on the contrary is all that he ever was, he is all that, in consideration of his primitive organization, he ever could have been.

¹ The term "idiocy" was at that time and for many years thereafter applied to all grades of mental defect.

Esquirol also called attention to the fact that mental deficiency is not a discrete category, clearly separated from the normal and manifested in all-or-none fashion. There are various grades of mental defect, separated from each other by countless imperceptible steps. "Who," he asked, "can recognize and describe all the nuances of mental condition that separate the thinking man from the idiot who lacks even the basic instincts?" Nevertheless, he considered that for practical purposes two classes of mental defectives may be distinguished, the merely "weak-minded" and the idiots, properly speaking. In the first class belong those whose mental and physical organization is more nearly complete: "They have understanding, ideas, memory, griefs and desires, but all of these are weak." The idiots, on the contrary, "have but few ideas; their emotional life centers about their bodily needs, they speak, at the most, only single words or syllables or express themselves only by inarticulate cries, they are lacking in sensitivity, observation and memory and are little or not at all amenable to training." A number of case histories which illustrate the differences between the two classes of defectives and also show how markedly each departs from the normal standard are presented.

Esquirol was very conscious of the need for some objective criterion by means of which the feeble-minded could be clearly distinguished from the normal, and the various grades of mental defect might be classified. Like others of his day, he clung to the hope that physical measurements, especially measurements of the size and proportions of the skull might yield an answer to his problem. Data on head measurements are presented for each of the individual cases which he reports, together with statistics collected by other persons who had worked on the problem. It is interesting to note, however, that when he attempted to draw up a plan for a somewhat more finely graded system of classification, he abandoned the use of anthropometry in favor of a functional system. In so doing he voiced the same conclusion as was reached by Terman some fourscore years later.² Esquirol stated the principle as follows:

Since speech, this exclusive manifestation of humanity which to man alone is given for the expression of his thoughts, is the sign which bears the closest relation to the intellectual abilities, it is thus the distinguishing characteristic of the major varieties of idiocy.

In the highest level of mental defect, the speech is easy; in the next lower level the speech is difficult and the vocabulary is more limited.

² Terman, L. M., Kohs, S. C., *et al.* (1918). In comparing the value of the different items in the 1916 Stanford revision of the Binet scale for predicting total score on this scale, Terman and his co-workers found that by means of the vocabulary test alone an estimate of a child's mental age may be derived which, in the majority of cases, will fall within one mental year of that obtained by use of the total scale.

In the highest level of true idiocy, the idiot can use only single words and very short sentences.

Idiots of the second grade articulate only single syllables but they make strong outcries.

Finally, the idiot of the third grade has no speech at all, neither sentences nor words nor even single syllables.

Esquirol's chapter on idiocy represents the first systematic attempt to organize the available knowledge in the field of mental deficiency and by means of classification and definition of terms to correct misconceptions and provide at least the beginning of a scientific vocabulary. Although he was by no means the first to point out the congenital nature of mental defect, an idea which had been either implied or vaguely stated in the writings of such men as Sauvages, Segar, Vogel, Linnaeus, Cullen, and others for half a century or more previous to the publication of Esquirol's monograph, up to that time the character of the behavior of the afflicted person rather than its point of origin had nevertheless been made the primary basis for the distinction between mental disorder and mental defectiveness. The schizophrenic who behaved like an idiot was likely to be classed as an idiot, even though he had lived the life of a normal person for a quarter of a century before the onset of his disease. Moreover, previous to the time of Esquirol, few scientists had taken more than a cursory interest in the problem. Early textbooks on mental disorders frequently made no mention of idiocy; at most the discussion was limited to two or three brief paragraphs. Esquirol's chapter is thus outstanding for its length and completeness as well as for its vigorous presentation of his point of view.³

AMERICAN INTEREST IN MENTAL DEFECT

Two events occurring near the middle of the nineteenth century gave tremendous impetus to the study of mental deviations in the United States. The first was the establishment of the earliest American scientific journal dealing exclusively with this topic—the *American Journal of Insanity* founded in 1844. Although, as the name implies, its major interest lay in the field of mental disease, the earlier volumes of the *Journal* also include a number of articles on the mentally subnormal. The creation of a recognized professional organ for the presentation of scientific papers in this area did much to awaken an interest in the field and called attention to the need for research on the many controversial

³ I have not had direct access to the original monograph but only to the German translation by W. Bernard which is presumably accurate and unabridged. In this edition the chapter on idiocy is 51 pages in length—a sizable treatise for its date.

issues which immediately became apparent as opportunity for the airing of opinion before a wider audience was given.

The second event of importance for the advancement of interest in mental differences occurred in 1848 when, because of political difficulties, Dr. Edward Seguin decided to leave France and make his home in America. Seguin's coming was important because of the change in the general attitude toward the feeble-minded that he helped to bring about. Up to this time Esquirol's view as to the incurability of mental defect was accepted rather literally both by the medical profession and by the general public. An unfortunate corollary of this belief was the feeling that if there was no hope of curing the condition, then it would be a waste of time to attempt to do anything at all in the way of improving the behavior of the unfortunate persons so afflicted. This view is well exemplified in the words which Itard addressed to the famous Wild Boy of Aveyron⁴ in a moment of discouragement over the child's slow progress:

Unfortunate one! Since my pains are lost and my efforts fruitless, take yourself back to your forest and primitive tastes; or if your new wants make you dependent on society, suffer the penalty of being useless and go to Bicêtre there to die in wretchedness.

But Itard's bitter disappointment over the outcome of his experiment was not shared by everyone. The French Academy of Science rightly looked upon it as a contribution of great scientific importance, the significance of which was by no means dependent upon the nature of the outcome. The academicians noted, moreover, that although his pupil had by no means attained normal intelligence, a tremendous improvement over his early condition had taken place. He had learned to speak one or two words, though not distinctly, and to recognize a number of printed words which he would match with the corresponding objects.

⁴ Near the close of the eighteenth century a naked boy, apparently about twelve years of age, was found wandering in the woods at Aveyron near Paris. The child was unable to speak, making only inarticulate, animal-like noises. He ran on all fours like an animal, drank water by lying on his face and sucking it up like a horse or ox, and when angered fought with teeth and nails. He was brought to Paris where he was examined by the famous psychiatrist Pinel, who pronounced him an idiot. Itard, a colleague and former student of Pinel, did not accept this verdict but believed the child's condition to be the result of isolation from birth or from an early age and consequent lack of opportunity to acquire the habits and skills of civilized man. He accordingly undertook the task of civilizing and educating the boy, but his efforts resulted in very meager success. Although Itard never formally admitted that Pinel's verdict was correct, after five years of painstaking effort had failed to bring the boy to a point where it was even remotely likely that he would be able to maintain himself in an ordinary society, Itard decided to discontinue his efforts at further training. The boy was placed in charge of a caretaker with whom he remained until his death.

He understood simple commands and had formed a fairly large number of associative responses to the routine of his daily living. Although his personal habits were still rude, they had undergone a measurable change in the direction of those common to civilized persons.⁵

Of many who became interested in Itard's study, perhaps no one was more quick to appreciate its fundamental significance than his gifted pupil Seguin. Born in 1812, Seguin early became interested in the question of mental defect and its possible amelioration. Although he studied with both Itard and Esquirol, he seems never to have fully accepted their point of view as to the incurability of mental deficiency. Certainly he was more impressed by Itard's account of the progress made by his savage pupil than by the failure of the latter to attain normal standards. Throughout his life he was an ardent protagonist of the doctrine that the feeble-minded can be educated, and that the success of the educator cannot be judged by any absolute standard of results to be attained but only in terms of the amount of improvement made by the learner. In reviewing the work of Itard he quoted with approval the official report by Dacier, the Permanent Secretary of the French Academy, which ran as follows:

The Academy, moreover, cannot see without astonishment how he [Itard] could succeed as far as he did, and thinks that to be just toward M. Itard, and to appreciate the real worth of his labors, the pupil ought to be compared only with himself; we should remember what he was when placed in the hands of this physician, see what he is now, and more, consider the difference separating the starting point from that which he has reached. [Seguin, 1907 edition, p. 21.]

The influence of Seguin in changing the attitude of the scientific world and that of the general public was profound. According to the pessimistic view then current, if mental defect cannot be cured it can only be endured, and all attempts at educating its unfortunate victims are wasted if they do not result in a genuine correction of the defect. With the unquenchable optimism and unflagging zeal of the born crusader, he proclaimed his faith that through suitable training even the idiot may be brought nearer to normality while many of those less severely affected can be made partially or wholly self-supporting. That this was not merely an idle claim was demonstrated by the results obtained in his own teaching. In 1837, at the early age of twenty-five, he established what was probably the first really successful school devoted entirely to the education of mentally defective children. Visitors to the

⁵ Itard's account of his experiment is embodied in two pamphlets, *De l'éducation d'un homme sauvage* (1801) and *Rapport sur le sauvage de l'Aveyron* (1807). An English translation by George and Muriel Humphrey, entitled *The wild boy of Aveyron*, is published by Appleton-Century-Crofts Company, New York.

school were impressed by the changes in the general manner and appearance of the children as well as by their proficiency in simple manual arts and even in the rudiments of reading and writing. Scientists and educators from all parts of the world came to wonder and admire. In 1842 Seguin was invited to take charge of the division for the care and training of the feeble-minded at the Bicêtre, but he found the rules and practices of so large an institution too restrictive for him to carry out his ideas as he wished. After a year at the Bicêtre he re-established his own school at the Hospice des Incurables, where he remained until his emigration to America in 1848.

Not only through his teaching and public lectures but even more through his writing, the ideas of Seguin gained wide recognition. The French Academy publicly commended him for his work and it is said that after reading one of his books, Pope Pius IX in a personal letter warmly praised him for the service he had rendered humanity. (Wallin, 1921.) Gone were the days when the feeble-minded were looked upon as hopeless burdens, incapable of training. A new optimism was springing up, a belief that by the right educational methods begun at a sufficiently early age many, if not all, might be made capable of leading at least an approximately normal life in society.

The spirit of democratic idealism in the United States provided an ideal soil for the implantation of this idea. Up to the middle of the nineteenth century, no systematic plan for the care and training of the feeble-minded existed in this country. If no other means of support was available, they were commonly looked after in workhouses, poorhouses, or similar public institutions. A fair number found their way into prisons and reformatories, and some of the lower-grade cases, especially if they were excitable and hard to manage, were placed in mental hospitals, where they were likely to be kept under more or less rigid physical restraint. But reports of the work of Seguin in Paris, as well as that of Guggenbühl, who in 1842 had established a school for cretins at Abdenburg, Switzerland, and that of White's school for idiots founded at Bath, England, in 1846, were not long in reaching the ears and arousing the interest of forward-looking persons in the United States. In 1846, Judge Byington of Massachusetts brought the matter to the attention of the state legislature, with the result that on April 11 of that year an act was passed which authorized the appointment of "Commissioners to inquire into the condition of the idiots of the Commonwealth, to ascertain their number and whether anything can be done in their behalf."⁶ The commission was appointed with Dr. S. G. Howe, at that time proba-

⁶ See "The report of the Massachusetts provisions for the insane and idiotic; at present and in olden times." *American Journal of Insanity*, 1848, 5, 371-377.

bly the best-known American authority on mental defects, as its chairman. During the next two years the commission carried out a complete survey of the State of Massachusetts with the aim of locating all cases of mental defect within its borders. In 1848 their findings were reported to the legislature as follows:

We have considered [as idiots] all persons whose understanding is undeveloped, or developed only in a partial and very feeble degree or who have lost their understanding without becoming insane as proper subjects for examination. Of the 574 persons reported to us as idiotic, 420 may be considered as properly idiotic, for their feebleness of intellect is connate; while 154 have become idiotic after birth Of the 420 idiots proper, 19 can now earn their board and clothing under the management of discreet persons; 141 do earn their board when properly managed; 110 can do trifling work if carefully watched and directed; 73 are as helpless as children of seven years old; 43 are as helpless as children of two years old and 34 are as utterly helpless as infants.

After reporting that "other countries have shown us that idiots may be trained to habits of industry, cleanliness and self-respect," the commission, in recommending that steps be taken to provide some form of training for these afflicted persons, inquired: "Can the men of other countries do more than we? Shall we, who can transmute granite and ice into gold and silver and think it pleasant work—shall we shrink from the higher task of transforming brutish men back into human shape?"

The appeal was not without effect. That same year (1848) the legislature of Massachusetts appropriated the sum of \$2500 a year for a period of three years to be devoted to the experiment of teaching and training ten idiots in a special school to be established at South Boston under the direction of Dr. Howe, with Mr. J. B. Richards as teacher. Later the school was transferred to Waverley, Massachusetts, where it became one of the leading institutions of its kind in the United States.

The example of Massachusetts was soon followed by other states. In New York State an appeal to the legislature for better provision for the feeble-minded was made by Dr. F. P. Backus during the same year that the corresponding request was made in Massachusetts. In 1849 the first New York State institution for the care of these persons was opened on Randall's Island. The movement spread rapidly, both in the United States and abroad. Before 1870, no fewer than eighty public and private schools for mentally defective children and adults were opened in the United States, the European countries, Canada, and Australia. (Johnson, 1894.)

Two problems of both practical and scientific import confronted the persons engaged in this work. First there was the question of selecting

cases for admission to the schools. Then as now, the number of applicants vastly outnumbered the places available within the schools, and community pressure was greatest in those cases where misconduct as well as mental defect made the child an unwelcome member of society. It thus became necessary for those in charge of admissions to be constantly on their guard in order to avoid turning their schools into mere reformatories for delinquents. That not all feeble-minded persons are of the same degree of defectiveness had long been recognized, and some idea of finding an objective basis for identifying those who should be classed as "idiots" (that is, mentally defective) was apparent even before the nineteenth century in the old English law code which stated: "It is sufficient to find him so [viz., idiotic] if he has not any use of reason, as if he cannot count twenty pence, if he has not understanding to tell his age or who is his father or mother." In Howe's report to the Massachusetts Legislature quoted in an earlier paragraph, a crude form of classification was used in which, it is interesting to note, the general concept of mental age was employed, although we are not told how the classification was determined. Since Howe apparently interviewed the cases himself, it is probable that some general system of questioning was employed. It is unfortunate that no record of the procedure seems to have been kept. The reference to age as a standard for describing mental level was apparently not uncommon even as early as the middle of the century. An example is found in a report (Hall, 1848) of the trial of a Negro named William Freeman, who had wantonly murdered four persons, apparently as the result of paranoid delusions which had their origin in a five-year imprisonment for a crime of which he was subsequently found to be innocent. During his incarceration he had been very brutally treated. At one time he was struck on the head with a wooden plank so violently that the plank was splintered. A serious concussion of the brain and subsequent deafness resulted from the injury. The report given at his trial clearly indicated that from the time of his release some degree of mental abnormality was present though the testimony indicated that up to the time of his imprisonment he had been mentally normal. The report of the medical experts shows the confused state of scientific thinking on this topic. Much importance was attached to the results of a phrenological analysis. The concept of age as a standard of reference may be noted in the testimony of one of the medical experts, a certain Dr. Dimon, who was asked to state his impression of the prisoner's mental competence. Dr. Dimon testified:

I should not think he has as much intellect as an ordinary child of fourteen years. In some respects he would hardly compare with children of two or three years.

Questions from the judge: With a child of what age would you then compare him in respect to knowledge?

Answer: With a child of two or three years old.

No facts are given to substantiate this opinion apart from the prisoner's usual response of "I don't know" to practically all questions asked him. Although it is stated that he was by this time almost totally deaf, there is no evidence of any attempt to ascertain whether or not he was able to hear and understand the questions put to him.

But the time was not yet ripe for systematic use of the developmental progress of the normal child as a standard of reference for judging the mental level of those whose minds had not developed normally. Not until sixty years of groping for signs and symptoms had elapsed—sixty years of patient measuring of skulls and other anatomical features, sixty years of hopeful experimentation with tests of sensory acuity, motor speed, and accuracy, and, with slightly greater success, with tests of rote memory and discrimination—was the amazingly simple and effective concept of "mental age" to spring to life in the versatile mind of Alfred Binet. Meantime, however, with the multiplication of institutions for the mentally defective both in the United States and abroad, the need for some practical system of diagnosis and classification became increasingly acute. At the same time, more general recognition of mental defect as a condition for which the individual is not himself responsible, and which renders him less capable of exercising intelligent control over his own acts, gave rise to public interest and scientific inquiry into another age-old problem—that of the juvenile delinquent.

THE YOUNG OFFENDER

Up to the early part of the nineteenth century, treatment of the child offender was generally both brutal and unintelligent. Punishments commonly imposed even for minor offenses were the whipping post, imprisonment for long periods, or even death. According to the General Laws adopted by the colony of New Plymouth in 1671, continued disobedience or rebellion against parental authority in a child of sixteen years was accorded the death penalty. (Abbott, 1938.) Although under the common law both in England and in the colonies, a child was not assumed to be morally or legally responsible for his acts before the age of seven years, children of no more than eight or nine years might be, and sometimes were, adjudicated serious criminals and made to suffer the full penalty for their acts. Nevertheless, even as early as the latter part of the eighteenth century, there was formal recognition of individual differences in intelligence as manifested by the ability to distinguish

between right and wrong. According to a quotation from Blackstone (1795) cited by Grace Abbott,⁷

. . . the capacity for doing ill or contracting guilt is not so much measured by years and days as by the strength of the delinquent's understanding and judgment. For one lad of eleven years old may have as much cunning as another of fourteen; and in these cases our maxim is that *malitia supplet aetatem*.⁸ Under seven years of age, indeed, an infant cannot be guilty of felony; for then a felonious discretion is almost an impossibility in nature; but at eight years old he may be guilty of felony. Also under fourteen, though an infant shall be *prima facie*⁹ adjudged to be *doli incapax*,¹⁰ yet if it appear to the court and the jury that he was *doli capax*¹¹ and could discern between good and evil, he may be convicted and suffer death. Thus a girl of thirteen has been burnt for killing her mistress; and one boy of ten, and another of nine years old, who had killed their companions have been sentenced to death, and he of ten years actually hanged; because it appeared upon their trials that the one hid himself, and the other hid the body he had killed, which hiding manifested a consciousness of guilt and a discretion to discern between good and evil.

Thus it would appear that in England at that time a child's very life might depend upon the mental rating assigned to him by a judge whose knowledge of child behavior may well be questioned. Certainly a more accurate method of mental appraisal was urgently needed.

Although with the passage of time the treatment of juvenile offenders became somewhat more lenient, and public sentiment both in the United States and abroad became increasingly opposed to the use of capital punishment for children, the question of individual responsibility remained a problem for the courts. Gradually, as the concept of insanity as a disease rather than as an indication of demonic possession or of divine wrath emerged from the fogs of superstition by which it had been surrounded, and as the distinction between mental deficiency and mental disorder became more clearly understood, the practice of calling in medical experts to testify as to the mental competence of accused persons became more common.¹² But then, as now, the experts frequently were unable to agree.

The question of the diagnosis of mental capacity was regarded by most people as solely a medicolegal problem till well on to the close of the nineteenth century. Toward the end of that century, however, a new influence began to make itself felt. Experimental psychology, up to then

⁷ *The child and the state*, II, 342

⁸ Evil intent is more important than age.

⁹ At first view.

¹⁰ Incapable of guilt, that is, not mentally responsible for his acts.

¹¹ Capable of guilt, that is, mentally responsible for his acts.

¹² See the account of the trial of William Freeman (pp. 10-11).

largely dominated by men like Weber, Fechner, and Wundt, who were mainly concerned with the study of sensation and its attributes, was extending its researches into wider fields. In England, Sir Francis Galton was conducting his famous investigations of mental inheritance, of sex differences in mental traits, and of the characteristics of famous men. In France, Alfred Binet and his associates had already embarked upon the long series of experiments on the mental characteristics of children which led, eventually, to the development of the first really successful method of classifying individuals with respect to their mental capacities. And in America, Cattell and others had succeeded in arousing enough interest in the problems of mental measurement among the members of the newly founded American Psychological Association to lead to the appointment of a special committee to study the matter and to formulate, if possible, a series of "mental tests" for use in the classification and guidance of college students. All this was a far cry from the reliance upon such matters as physical appearance, the shape of the head, the "glance of the eye," or the unsubstantiated reports of poorly qualified judges—criteria which, up to then, had been the only available bases for diagnosis and classification. The early tests, as will be shown in a later chapter, were far from perfect. Many of those which were first tried proved to be even less effective than the subjective judgments of physicians and teachers. Nevertheless, they constituted a step in the right direction since they called attention to the necessity of utilizing a standard situation, of providing a common basis of reference, if individuals are to be classified in a uniform and meaningful way. The standards of different people vary too greatly to be useful measuring rods.

The Educational Need

In 1908, a report by E. L. Thorndike published by the United States Bureau of Education became front-page news for American readers. This report presented statistics on school enrollment for many of the leading cities of the country. By comparing age-grade proportions at the various levels, Thorndike arrived at the following widely quoted conclusion:

I estimate that the general tendency of American cities of 25,000 and over is, or was at about 1900, to keep in school out of 100 entering pupils 90 till Grade 4, 81 till Grade 5, 68 till Grade 6, 54 till Grade 7 and 40 till the last grammar grade.

According to Thorndike, in 1900, fewer than 10 per cent of the children entering first grade ever completed high school.

But this, as Thorndike pointed out, is not to be interpreted as meaning that a large percentage of children ceased to attend school while still at a tender age. They remained in school but failed to progress through the grades. Some, to be sure, among the many who were made discouraged and rebellious through repeated failure, contrived to evade the law before the period of compulsory school attendance had expired. Although as early as 1900 most of the more progressive states had passed laws requiring attendance of all children¹ under the age of thirteen or

¹ The first state to pass a compulsory attendance law was Massachusetts in 1852. This law was not clearly phrased and was amended in 1859 to require annual attendance of all children between the ages of 8 and 14 years for at least 12 weeks, of which 6 weeks were to be consecutive. However, exceptions were permitted if parents were too poor to send the child to school or "if his bodily or mental condition was such as to prevent his attendance." The law was poorly enforced for many years and was repeatedly amended so as to make evasion more difficult. Satisfactory enforcement was not attained until about 1890. By this date, 32 other states had enacted attendance laws but in many of them enforcement was still lax. As late as 1910, the United States Bureau of Education reported that of children between the ages of 7 and 13 years, only about 86 per cent were attending school, and of those 14-15 years, 75 per cent were in attendance. In 1914, 6 states still had no compulsory education laws, and 3 others did not require attendance beyond the age of 12 years. In 10 other states, the upper requirement was 14 years, and only a single state had set a limit higher than 16 years. Moreover, in many states there were a sufficient number of loopholes in the law to permit some children to drop out before the official age had been reached.

fourteen years, these laws were often poorly enforced. As the upper limit of compulsory attendance was gradually extended, in many states both age and school accomplishment were taken into account in fixing the time at which a child might legally drop out of school. By 1910 a fairly common requirement was that children should remain in school until the age of sixteen unless they had finished a given grade, usually the fifth or the sixth, in which case they might leave at some specified younger age.

The intention of such a law is obviously that of ensuring that no child shall leave school until he has secured at least the minimal essentials of an education. But its weakness lies in the assumption that merely attending school for a greater number of years means better education. This assumption unquestionably holds good for the average of groups but not necessarily for all the individuals comprised within the groups. With the type of educational system followed in most schools at around the turn of the century, the age-grade specification for the upper limit of compulsory education was likely to be least effective for the very children whom it was designed to benefit. For a child who has not succeeded in getting as far as the fifth or sixth grade in school by the age of fourteen, will not, in most instances, profit greatly from two or more additional years of failure.

And continued failure, while not the invariable consequence of requiring the laggards to remain in school after their more able fellows had been permitted to leave,² was dismayingly common. It thus came about that the more rigid enforcement of compulsory education laws, together with the upward extension of the age limits set by these laws which took place at about the beginning of the present century, was not wholly an unmixed blessing. Certainly it was a good thing for most children and greatly improved the general educational level of the population as a whole. But it also resulted in the retention in the elementary schools of many children who learned slowly or not at all by the educational methods then employed.

In the first decade of the twentieth century, there was little concept of adjusting the curriculum to the needs and abilities of the individual child. Grade requirements were rigid and formalized. Promotion de-

² It is hardly necessary to say that the greater number of the more able children remained in school beyond the age at which the law permitted them to leave. Nevertheless, cases were far from rare in which ignorance or cupidity on the part of parents led to the withdrawal of children who had met the educational requirement at the lower age limit and who would have profited by further school attendance, while those less able to benefit by it were forced to remain in school. Thus, those who most needed legal protection of their rights to an education were denied it; while those for whom such a law was of little advantage were provided for.

pended upon a child's mastery of the subject matter prescribed for his grade. Teachers were expected to know the subjects which they taught and to be able to enforce order in the classroom. That they should also know something about children was a novel idea, just beginning to affect the thinking of a small number of the more advanced educators. The backward child, the slow-learning child, was known to every teacher at firsthand, but few had any idea of dealing with him beyond nonpromotion and punishment if he failed to conform to classroom rules. Thus it came about that the primary grades in many cities, especially those having many children of foreign-born parents, became clogged with children whose ages ranged all the way from six to sixteen years. Rarely were any special arrangements available for the older children. In many cases, not even desks of suitable size were provided. Long-legged adolescents of fourteen and fifteen years were forced to crowd themselves into seats of a height suitable for six-year-olds. Year after monotonous year they were expected to struggle with the intricacies of the same primer, to cramp their clumsy fingers in laborious formation of the letters of a meaningless alphabet, repeat on demand such unintelligible formulas as *c-a-t*, *cat*. Small wonder that they often rebelled. Small wonder that teachers became exasperated with them. The problem was real and serious.

Some idea of the extent of retardation then common is given by Table 1, which is based upon the Report of the Ohio School Survey Commission of the city of Dayton for the school year of 1912-1913.

TABLE 1
AGE-GRADE DISTRIBUTION
OF 11,769 CHILDREN IN THE ELEMENTARY SCHOOLS OF DAYTON, OHIO
DURING THE SCHOOL YEAR OF 1912-1913*

Grade	Age													Per Cent 3 or More years Overage
	5	6	7	8	9	10	11	12	13	14	15	16	17+	
I	309	1308	451	128	52	11	6	4	2	1	3.3
II	..	203	906	372	115	52	30	8	..	2	5.5
III	188	823	455	169	73	45	19	2	2	7.9
IV	168	621	414	224	113	43	19	2	2	..	11.1
V	2	179	475	307	236	111	51	6	1	..	12.3
VI	4	148	439	344	166	69	16	5	1	7.6
VII	8	158	430	318	159	26	5	..	2.8
VIII	8	104	334	221	84	10	2	1.6

* Ohio State School Survey Commission. *Overage and progress in the public school of Dayton, Ohio*. Dayton, Ohio: Bureau of Municipal Research, 1914.

The first grade includes children who range in age from five to fourteen years. Of these children it is stated that 128 had already spent not less than two and a half years in that grade, that is, they were entering upon their third year or more of first-grade work, while in the entire city there were 330 children who had been enrolled in their present grade for the third (or more) consecutive year.

The last column of Table 1 shows the percentage of children in each school grade who were three or more years overage for that grade. As failures accumulate, the percentage of children retarded three or more years steadily increases in each grade up to the fifth, where approximately one child in eight had presumably failed to make his annual promotion at least three times. The tendency for the backward children to drop out of school as soon as the law permitted is clearly shown in the decreasing number of children three or more years overage for their grade in the grades beyond the fifth.³

TABLE 2
NUMBER AND PERCENTAGE OF CHILDREN RETARDED IN SCHOOL BY
THREE OR MORE YEARS IN CERTAIN AMERICAN CITIES IN
DECEMBER, 1908 AS REPORTED BY THE U.S. BUREAU OF EDUCATION

City	School Population 9 Years or Older		Number Retarded 3 or More Years		Per Cent Retarded 3 or More Years	
	Boys	Girls	Boys	Girls	Boys	Girls
Birmingham, Ala.	2,534	2,916	632	634	24.9	21.7
Detroit, Mich.	14,000	13,558	1,096	747	7.8	5.5
Los Angeles, Calif.	11,151	11,164	1,007	609	9.0	5.5
Philadelphia, Pa.	47,859	47,619	5,860	4,999	12.2	10.5
Providence, R.I.	8,814	8,928	768	639	8.7	7.2

The Dayton conditions were by no means unique. An equal or greater amount of retardation was found in most cities and towns throughout the country, while in the rural schools the situation was likely to be even worse. A report by Strayer (1911) gives the results of a survey made by the United States Bureau of Education in December, 1908. Although this report does not present age-grade distributions in detail, it does show, for a large number of towns and cities in all parts of the United States, the number of children retarded in grade placement by varying amounts up to five years and above. Table 2 shows the number of children who were retarded three years or more in each of five cities in different parts of the country.

³ The minimal age in Ohio for securing working papers at the time this census was taken was 14 years. The following year it was raised to 16 years.

While retardation was then, as it is today, much more common in the South than in the other parts of the country, the number of children everywhere whose school life was merely a series of disheartening failures was disturbingly great. By the early part of the century, moreover, it was becoming apparent that enforcement of the compulsory education laws could not solve the problem of school retardation. The old saying, "One man can lead a horse to water but twenty thousand cannot make him drink," once more proved its truth. In spite of long and regular attendance, some children could not and did not learn by the ordinary methods of classroom instruction. Not only was the time spent in school unprofitable to themselves but in many cases such children had a disrupting effect upon schoolroom discipline. Often so much of the teacher's time and energy would be consumed by the backward children in her group that little was left for teaching those who were better able to learn.

Taxpayers as well as teachers and school administrators gradually became aware of the problem. Statistics on the cost of school retardation received wide publicity. It was shown that millions of dollars were spent annually in a vain attempt to force upon all children a kind of education that many were unable to acquire. Slowly but surely both schoolmen and the public in general began to realize that failure to learn is not always due to laziness or depravity on the part of the child and that something other than punishment is needed in order to solve the problem. A different kind of curriculum as well as different methods of teaching is called for.

Here and there a few cautious experiments were tried. As early as 1859, a special class for children who had been unable to do the work of the regular grades was established in Halle, Germany, in the hope that a period of special tutoring might suffice to enable these children to make up their deficiencies and return to their regular classes. The lack of success of this plan led to its abandonment within a few years. In America the first special class for retarded children was established in 1893 by the public schools of Cleveland, Ohio. Other cities soon followed suit. Interest in these classes was increased by the establishment in 1896 of a four-week summer clinic for the study of backward children at the University of Pennsylvania, under the direction of Lightner Witmer. The idea that the slow rate of mental development of the mentally retarded might provide a highly favorable opportunity for the study of normal development appealed to psychologists and educators alike. The normal child was compared to an express train running at full speed. His mental processes change and develop so rapidly that you can scarcely follow them. All you get is a blurred impression. The feeble-minded child, in contrast, was said to be like a slow freight, moving along the same track

as that followed by the express but at a pace that gives leisure for observation and study. "The first step," said Witmer, "toward the understanding and adequate training of normal and gifted children in the public schools is to understand the problem of individual training with backward and mentally defective children." (Witmer, 1911.)

Notwithstanding all this, the spread of the movement toward the establishment of special classes in the public schools was hampered by several factors. First, there was confusion as to the major purpose these classes should endeavor to serve. Should their first aim be to ameliorate the condition of the children either by special tutoring with the hope of eventual return to regular classes or by special industrial training designed to prepare them for some kind of useful work after leaving school? Or should the major emphasis be placed upon the relief given to the regular classes through withdrawal of the laggards? In theory, at least, the concept of equal educational opportunity for all led to the general acceptance of the idea of aid to the handicapped child as the major objective of the special class, but in the practical situation, the second of the two motives was likely to operate. It was but natural that teachers and school principals should urge the selection for placement in the special classes of those children who were giving most trouble in the regular grades. Had teachers been given a free hand, these classes would soon have become mere disciplinary rooms for juvenile delinquents. An outside criterion for selection was evidently needed, but there was no general agreement as to what that criterion should be.

Problems arising from individual differences in ability to learn were not confined to the public schools. College enrollment was increasing, and among those who sought college training there were many who proved unable to do the required work. Moreover, the number of scholarships and prizes available for the more promising students was increasing, and gifts of this kind were more likely to be made available if an institution could point to previous recipients of outstanding accomplishment. The selection and classification of college students thus became a question of real concern to forward-looking members of college faculties, even before the opening of the present century. Of this we shall have more to say in a later chapter.

The Scientific Background

EARLY ATTEMPTS AT DESCRIBING INTELLIGENCE

The fact that man is essentially a thinking animal has led philosophers in all ages to turn their attention to the nature of thought and the processes of thinking. With the rise of the so-called "faculty psychology," the use of the term "intelligence" to designate the successful apprehension of facts and their relationships became a general practice. Two well-nigh inevitable results of this were, first, the tendency to reify intelligence as something possessed by man and used by him in the acquisition of knowledge and the mastery of the physical world, and, second, the construction of an artificial framework or pattern to which all intellectual processes were presumed to conform. At the bottom of the structure were placed the *sensations* which, through repetition, combination, and association, gave rise to *perceptions*. Perceptions, in like manner, were said to become organized and generalized into *concepts*. The work of the intellect was thought to consist essentially in the transformation of mental events of a lower order to those of a higher order—in building perceptions out of sensations and concepts from percepts. This was believed to be accomplished by the operation of such "faculties" or facilitating mechanisms as attention, memory, judgment, and so on.

As far as its direct effect upon the testing movement is concerned, the contribution of the philosophers of the eighteenth and nineteenth centuries to the study of mental development was small. Indirectly, however, it played a not inconsiderable part. At a time when the exciting discoveries of the psychophysicists and the sensory psychologists threatened to indoctrinate all those interested in the study of mental processes with a molecular rather than a molar point of view, the philosophical controversies and pronouncements about the nature of *mind* provided a wholesome counterbalance. In the second place, artificial as they undoubtedly were, these theories nevertheless provided an initial framework, a series of working hypotheses regarding the kind of activi-

ties likely to be most significant as indicators of intelligence which Binet and his collaborators could put to actual trial later on.

Bain's *The senses and the intellect*, published in 1855, and Taine's two volumes, *On intelligence*, which appeared in 1870, helped to lay the groundwork for the idea that mental activity is not merely a subjective affair, knowable only to the individual, but has its external signs. Taine, in particular, placed much emphasis upon the importance of the formation and utilization of abstract ideas as the essential basis of intelligent action. The early theories of associationism as propounded by such men as Locke, Hume, and Berkeley were modified by Taine into a more precisely defined scheme which included the processes of *substitution* (of one idea for another which thus becomes the sign or symbol of the first), of *elimination* (circumstantial or unstable features not common to all occurrences of the phenomenon are gradually dropped out), and finally *reduction* of the sign to its simplest and most general terms.

In this process [says Taine], in which the first term has taken on the property of standing wholly or partially in place of the second, so as to acquire either a definite set of its properties or all of those properties combined, we have, I think, the first germ of the higher operations which make up man's intelligence.¹

THE RISE OF ABNORMAL PSYCHOLOGY

Taine was also one of the first to see clearly how much the study of abnormal reactions can contribute to an understanding of the manner in which associations and reactions are built up in normal people. A considerable portion of the first volume of *On intelligence* is devoted to this idea, which was to become the keynote of French psychology during the last quarter of the century. He described in detail the origin and development of systematized delusions in a number of mentally diseased patients and stated that these phenomena "throw great light on the mechanisms of the mind."

The seed fell on fertile soil. The new psychology, a science of experiment and controlled observation, of hypotheses and their verification by facts secured at first hand, was hardly a decade old. Already, however, the cleavage which exists to this day between the "clinical" or the "social" psychologists and those whose interests center about problems with less immediate application to human affairs was beginning to manifest itself. In Germany the great laboratories of experimental psychology headed by that of Wundt at Leipzig adhered closely to the study of sensory phenomena, in which facts and principles were sought without regard to their immediate applications. In France, psychological interest from

¹ *On intelligence*, I, 6.

the beginning was largely concerned with the study of individuals who deviated from the normal pattern—the insane and the feeble-minded. In the beginning, the abnormal were studied chiefly for their own sake. Pinel, Esquirol, Itard, and Seguin were interested in learning what they could about mental disease and mental deficiency as such; they wished to ascertain the causes of such conditions, their course of development, and most of all, what could be done to alleviate them. But with the dawn of the era of scientific observation and experiment, as psychology cast off the leading strings by which it had been bound to its parent, philosophy, and began to step forward along a path of its own choosing, the behavior of the abnormal was seen from a new angle. Richet and Charcot, by their use of hypnosis as a therapeutic measure, did much to bridge the gap between normal and abnormal behavior. The genuineness of hypnotic phenomena and the possibility of inducing hypnosis in apparently normal subjects had been generally recognized since the days of Braid. Charcot demonstrated the greater susceptibility to hypnosis of certain types of mentally abnormal subjects and thus raised the question of hypnotic suggestibility as a possible symptom of the hysterical constitution.

Charcot's influence upon psychology in France was profound. Both Binet and Freud studied with him. His work with hysterical subjects at the Salpêtrière was continued by his brilliant pupil, Pierre Janet, who, through his lectures at the Sorbonne and as visiting professor at Harvard² was able to show, even more clearly than Charcot had done, how much an understanding of abnormal behavior can illuminate and clarify many of the puzzling aspects of the so-called "normal" personality.

Another French psychologist who saw in the abnormal mind the possibility of finding cues to normal mental processes was Théodule Ribot, who began his psychological career by making a systematic study of English and German psychological systems. The results of these studies were published under the titles of *La Psychologie anglaise contemporaine*³ (1870) and *La Psychologie allemande contemporaine*⁴ (1879). He thus brought to his later studies of mental deviation a thorough acquaintance with the psychological theories and experiments of three countries. From the outset his interest centered closely about the question of the origin and progress of mental characteristics. Genetics, as we know it, did not then exist as a science, but in England Galton had

² Janet's Harvard lectures were published in book form under the title *The major symptoms of hysteria*. 1907.

³ English translation entitled *English psychology*.

⁴ German translation entitled *German psychology of today: the empirical school*.

already published his *Hereditary genius* (1869) and it was probably this which stimulated Ribot to write his more general treatise *L'Hérédité psychologique*⁵ in 1873.

By the end of the seventies, Ribot's attention had become rather definitely centered upon the question of mental deviation in relation to the normal mind. At no time was he primarily concerned with mental abnormality as a condition to be remedied or with questions of prevention or treatment. His interest was psychological, not medical. His ever-recurring questions were: Is this behavior which we regard as abnormal really outside the range of normal responses, or does the difference lie chiefly in the conditions under which it appears? In his three monographs, *Les Maladies de la mémoire*⁶ (1881), *Les Maladies de la volonté*⁷ (1883), and *Les Maladies de la personnalité*⁸ (1885), he stressed the apparently great variety of ways in which a given mental function may become dissociated from its normal relationships, and the consequent necessity of seeking for some general principles under which the great diversity of signs and symptoms may be grouped. Eventually this led him to an ontogenetic approach in which both the deviations in behavior that fall within the range commonly regarded as "normal" and the more extreme divergences of definitely pathological cases were regarded as developmental phenomena resulting from environmental impact upon a given type of genetic constitution. This view is most clearly set forth in his *La Psychologie des sentiments*⁹ (1896). All this led rather directly to the early work of Binet, who published his *Psychologie de raisonnement*¹⁰ in 1886 and *Les Altérations de la personnalité*¹¹ in 1891.

MENTAL INHERITANCE IN MAN

Meantime in England Darwin's epoch-making book, *Expression of the emotions in man and animals* (1872) had appeared almost concurrently with the second edition of Herbert Spencer's *Principles of psychology* (1872). In both, the belief was expressed that the facial expressions and bodily postures commonly seen during strong emotion are not primarily the product of observation and imitation but are part of man's genetic inheritance from his prehuman ancestors. Thus not merely

⁵ English translation entitled *Heredity: a psychological study of its phenomena, laws, causes and consequences*.

⁶ English translation entitled *Diseases of memory; an essay on the positive psychology*.

⁷ English translation entitled *The diseases of the will*.

⁸ English translation entitled *The diseases of personality*.

⁹ English translation entitled *The psychology of the emotions* (1897).

¹⁰ *Psychology of reasoning*.

¹¹ *Changes in personality*.

his physical body but his behavior as well is to be regarded as a biological phenomenon, subject to the same laws and to be studied by the same methods as are used with other organisms. Moreover, if such apparently purposeless acts as frowning or uncovering the canine teeth in rage can be genetically determined, why may not special talents, defects, personality characteristics, or even differences in intelligence be likewise handed down from parent to child?

The idea of the evolutionary origin of human variation was, of course, not a new one. Early in the century it had been suggested by Erasmus Darwin, the grandfather of Charles Darwin. The latter had published his *Origin of species* in 1859. The first edition of Spencer's *Principles of psychology* (1855) includes a brief discussion of the evolutionary origin of emotional expression which attracted little general attention at the time though Darwin refers to it in his *Expression of the emotions*. Galton, whose attention had been caught by the observation that eminent men are likely to have eminent sons,¹² published his *Hereditary genius* in 1869, three years before the *Expression of the emotions*. This was followed by *English men of science; their nature and nurture* (1874), and several years later by *Natural inheritance* (1889). In a brief paper only six pages in length, which appeared in 1876 and was later reprinted in *Inquiries into the human faculty and its development* (1883), Galton also introduced a method of studying human inheritance which has since that time become widely used—the study of the resemblances of twins.

STATISTICAL METHODS

Galton's researches into mental inheritance in man were destined to have far-reaching influence upon scientific thinking. The questions which he raised almost three quarters of a century ago are still matters of active controversy, but his greatest contribution to mental testing as we know it today is not so much the subject matter of his investigations as the methods which he used to solve his problems. For Galton was the first man to see clearly that the only way of reducing the mass of chaotic impressions derived from observation of human beings to systematic order is through a quantitative approach. In 1846 Quetelet, a Belgian astronomer and statistician, had shown that the laws of probability originally developed by Gauss and Laplace can be applied to the physical measurements of man. He found that when such measures as the height

¹² It may be noted in this connection that both Charles Darwin and Francis Galton were grandsons of Erasmus Darwin. The Darwin family, as Galton notes in *Hereditary genius*, where he modestly refrains from mentioning his own name, includes a number of other eminent persons.

or weight of a relatively homogeneous group of persons (Quetelet used French soldiers as subjects) are arranged in order of magnitude along the abscissa of a curve with the number at each successive value plotted on the ordinate, the resultant curve corresponds very closely to what we now know as the *normal curve* or the *curve of probability*. Quetelet was greatly impressed by the fact that, in all of the measurements which he considered, the vast majority of individuals were found to stand at or near the average. Eventually he came to the conclusion that the average may be regarded as the ideal toward which nature is working. He believed that deviations from the average are in general to be regarded as departures from nature's ideal, however desirable or undesirable they may seem to us. Although Galton did not accept this interpretation, he was nevertheless greatly impressed by Quetelet's figures. More than by anything else, his imagination was captured by the fact that a given measurement can be expressed quantitatively, not only in terms of its own units of measurement (such as inches or pounds), but also in terms of the frequency with which it may be expected to occur in a given population. At once he realized the vast importance of this principle for the treatment of biological and psychological data. For it is obviously not the absolute measurement of a man or of a society which is of chief importance, but their relative standing within the group to which they belong. No matter how strong an army may be, it can be defeated by one still stronger. An Englishman of average height becomes a giant in a company of pygmies. Moreover, while we cannot compare height and weight directly because they are not expressed in the same units, we can do so with ease and accuracy if both are stated in terms of the extent to which they depart from the average when each is expressed in terms of its own variability. To us these ideas are commonplace; to Galton, who first envisioned their application to the study of human behavior, they were so exciting that the remainder of his long life was devoted to elaborating them. To Galton, also, we owe the concept of correlation, though the actual product-moment formula was the work of Pearson. As early as 1846, a French mathematician, A. Bravais, had worked out some of the basic theorems upon which the formula rests.

In 1904, Galton, then eighty-two years of age but still mentally vigorous, endowed a permanent fellowship for the study of human inheritance at the University of London. For this field of inquiry he had earlier proposed the name "eugenics."¹³ Convinced as he was that the only sound route to an understanding of human abilities is by way of mathematical analysis, he appointed the eminent mathematician Karl Pearson to be the

¹³ Eugenics. From the Greek *eu* meaning *well* or *good*, and *genus*, race or species, hence the study of methods for improving the racial stock.

first incumbent. Later on, this fellowship was merged with Pearson's own biometric laboratory, which had been established some years earlier than the Galton Fellowship.

It would be hard to overestimate the contribution of these two men to the development of the theory and methodology of mental measurement. Pearson was the younger by more than thirty years, but their

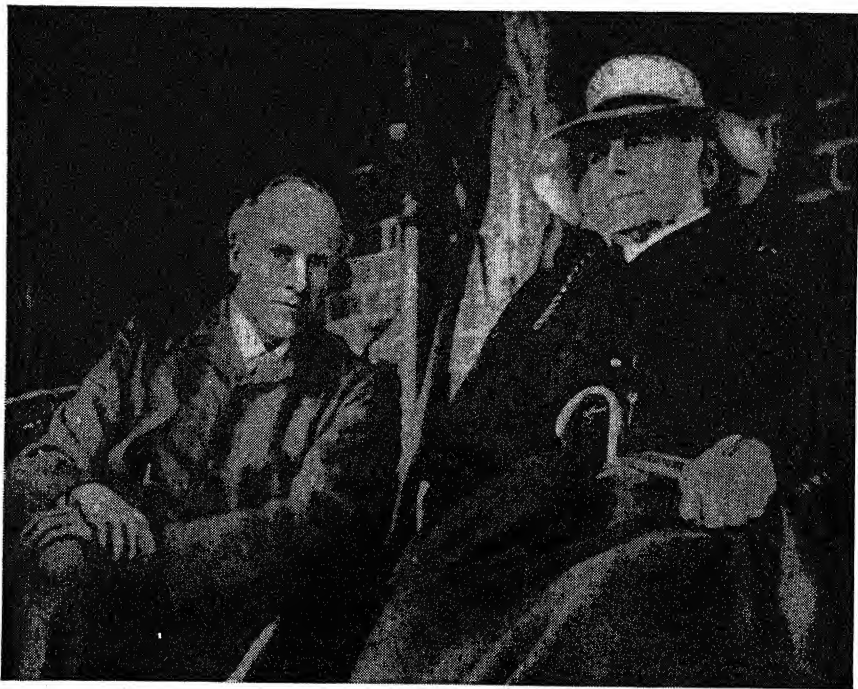


FIG. 1. FRANCIS GALTON, AGED 87, ON THE STOEP AT FOX HOLM, COBHAM, WITH KARL PEARSON, HIS BIOGRAPHER. (Pearson, left; Galton, right.) (Reproduced by permission of Professor E. S. Pearson from Plate XXXVI in Volume III A of the *Life of Francis Galton* by Karl Pearson. Photograph obtained through the kindness of Dr. Helen Walker.)

interests and professional outlook were similar. Yet each had his own method of working and each played his own unique part in giving form and direction to the new science. Galton, a man of wealth and leisure, was free to investigate every question that his brilliant and versatile mind propounded. Although for more than half a century his main concern was with questions of human inheritance, he could still take time off to ascertain the extent to which people differ in their ability to hear

ones of very high pitch or to recall in detail the appearance of their breakfast tables. For pastime he devised several clever schemes for getting a more realistic apprehension of the meaning of large numbers. However, in spite of his profound appreciation of the possibilities of mathematical analysis and his application of mathematical principles to the solution of problems in which he was interested, Galton was not, like Pearson, a highly skilled mathematician. His keen insight and intense intellectual curiosity which, once he had glimpsed an idea, would give him no rest until he had worked through to a solution of it, made him pre-eminently the pioneer who must always be pressing on to new horizons. He was a close observer of facts and their relationships, a bold thinker, fearless in the pursuit of his ideas. His was the inductive method. He noted a fact or a series of facts, and on that basis constructed a hypothesis which he then proceeded to test through the collection of a larger or more representative body of data. Galton was interested in statistical method as a tool to be used in the solution of concrete problems. Although he devised a number of statistical techniques, he did so with a concrete purpose in mind. Pearson, on the other hand, though he shared Galton's interest in eugenics, was more strongly attracted by abstract questions of scientific theory. He was interested in technical questions for their own sake as well as in reference to the problems which they were designed to solve. He liked compact mathematical demonstrations and well-organized theoretical expositions. His *Grammar of science* (1892) is still a classic in its field.

ADVANCES IN MEDICAL KNOWLEDGE

The early medical studies have only indirect bearing upon the methods of mental measurement which were to be developed in a future that was still remote. The Hippocratic oath, which is still taken by medical students, set a standard of scientific honesty and professional ethics that served as a lodestar throughout the dark ages of superstition and charlatanry that were to follow. Not until the eighteenth century, however, which Garrison (1929) has characterized as "the age of theories and systems," does a more nearly direct relationship appear. At this time Carl von Linné (Linnaeus), a Swedish botanist and physician whose great passion was for orderly classification, after first drawing up his famous system of classifying plants, tried to extend his method to the classification of animals, including man. After taking his medical degree he proceeded to devise a similar system for the orderly classification of diseases, including mental diseases.

The importance of the work of Linnaeus in the field of mental disorder does not lie in the particular system of classification which he

advocated and which now would seem both crude and fantastic. It is significant because it is a system. Although the principles upon which it was based were artificial, they were at least straightforward and internally consistent. Any such plan, whether it be right or wrong, has this great advantage over a haphazard collection of unorganized statements—it can be tested and evaluated as a whole. To Linnaeus, also, we are largely indebted for the rise of scientific appreciation of the importance of precise terminology and exact definitions.

William Cullen (1710–1790) of the University of Edinburgh did much to correct the current theories of “animal spirits” which had been commonly accepted since the days of Aristotle. The relation between thought and action—the mind-body problem—had puzzled the scientific world for centuries. For lack of a better explanation, the idea of a mysterious “essence” or nonmaterial “fluid” existing within the body and regulating its movements was generally accepted, a doctrine to which the strong religious sentiment of the times undoubtedly contributed. Albrecht von Haller (1708–1777) demonstrated that the functions of nerve and muscle differ, the former being characterized by “sensitivity,” the latter by “irritability.” Cullen regarded nerve and muscle as continuous, thereby explaining how one could affect the other. He taught that life itself is a function of nerve energy. For many years he headed the School of Medicine at the University of Edinburgh, where, both through his teaching and his writings, he exerted a profound influence upon scientific thinking. He was probably the first person in the medical world to assign to the mental diseases a rank of equal importance with other classes of disease in a systematic treatment of the ills of mankind. In his *Synopsis nosologicae medicae* (1785) he grouped all diseases into four great classes: (1) the febrile, such as typhoid fever; (2) neuroses, including epilepsy, mania, etc.; (3) diseases such as scurvy, resulting from bad diet or other habits or conditions affecting bodily health; and (4) localized diseases or conditions such as cancer, tumors, carbuncles, and the like.

Thus by the beginning of the nineteenth century, inching its way into the scientific world was the idea that mental phenomena are physiological rather than metaphysical, that mental pathology is the province of the physician, not that of the priest or the soothsayer, and that the way to better understanding of human behavior is to be found in controlled experiment rather than in the philosopher’s armchair. Not until another half century had elapsed, however, did the scientific movement as we know it today really find its stride. During the first half of the century there were some notable discoveries such as the demonstration by Magendie of the difference between motor and sensory nerves (1822),

an idea that had previously been suggested by Sir Charles Bell without wholly convincing proof. But with the rise of physiological psychology around the middle of the century, the era of scientific experiment in the biological sciences moved suddenly into high gear.

THE PSYCHOLOGY OF SENSATION AND ITS RELATION TO MENTAL TESTING

Experimental psychology in Germany during the latter half of the nineteenth century was chiefly concerned with the study of sensation and the sense organs. Although the discovery by astronomers of the important fact that two observers occupying equally favorable positions will rarely agree precisely as to the time an event is perceived might have turned the attention of psychologists to the study of individual differences in ability and behavior, it failed to do so. Not until many years later did the study of the individual person become an active topic for psychological research. Perhaps the fact that the discovery was made in connection with a field in which the interest was so completely centered upon the external event that the observer was regarded merely as an instrument was responsible for the slight interest which psychologists took in the matter. Perhaps the time for such an interest was not yet ripe. Certainly it was necessary to learn much more about the external factors affecting the behavior of persons in general before it could be safe to impute to the individual personality that which might well be the result of some uncontrolled factor in the external world.

It is often said that the work of the German laboratories at this time pointed in the opposite direction from that of Binet and his associates, who were, from the beginning, concerned mainly with the study of individuals. Superficially considered, there is unquestionably a great divergence between the two. In the German laboratories the basic question in most cases was: What is the effect of this experimentally introduced change in the conditions of observation upon the reported perceptions of the subjects? Differences in the reports of different people were in general looked upon as errors, resulting either from imperfect control of conditions or from carelessness on the part of the observers, and every effort was made to reduce these differences to a minimum. Nevertheless, the fact that such differences did not disappear, even after the most rigid controls had been used, interested Cattell to such an extent that in spite of Wundt's lack of sympathy with the problem, he chose the topic of individual differences for his Ph.D. thesis. Of Cattell's work along these lines more will be said in the following chapter. For the moment, however, we may note that had the question of individual differences been

differently regarded in Wundt's laboratory at the time Cattell was an assistant there, it is entirely possible that his attention might have been turned in some other direction. There is nothing like a long-continued effort to eliminate something for convincing one of its fundamental importance!

Thus in spite of the fact that the work of the German laboratories seems at first glance to be so diametrically opposed to the problems of mental measurement as we know them today, at least two discoveries of great importance for the mental testing of the future may in large measure be credited to them. First there were the many clear demonstrations of the marked effect upon the responses of the subjects of slight changes in the instructions given to them or in the manner in which the stimuli were presented. Changes in the appearance of colors when mounted on different backgrounds or when viewed under different conditions of illumination, the effect upon reaction time of changing from an auditory to a visual stimulus or of shifting the attention of the subject from the expected stimulus to the response to be made, as well as many other seemingly unimportant modifications in the conditions of experiment were found to be factors of no small consequence in determining the results. All this led to the early recognition of the need for clear and precise statements of the manner in which tests are to be given and scored and of the necessity for maintaining exact uniformity of procedure if interpretations are to be made on a uniform basis. Second, the fact that no means of external control that have yet been found will eliminate the differences in the performances of different subjects on the same task lent confirmation to the belief that even after all outside factors have been made uniform, there remains something to be measured. To us this is a truism; in the last quarter of the nineteenth century it was an idea to be challenged at least as far as the great mass of "normal" human beings is concerned. No one doubted that there are idiots and insane. Every teacher knew that children differ in their ability to learn. But in an age dominated by the idea of control by experimental variation of conditions, it was not at first easy to accept the view that the statistical study of variations among human beings, when the experimental conditions are kept the same, might be an equally valid and important field of scientific research. It was hard to eradicate the suspicion that the apparent differences among the subjects, the obstinate "personal equation," might, after all, be reduced to experimental control if all the correlated external facts could be ascertained. Possibly it never could have been disposed of except by the failure of its own proponents to do so after rigorous effort. Thus the work of the psychophysicists served to point the need for a *different kind* of psychological research by their

very failure to account for all the sources of variation in response. Not Binet only, nor Binet and Cattell, but Wundt himself, though all unknowingly, must be regarded as one of the progenitors of mental testing.

THE CHILD STUDY MOVEMENT

Locke's concept of the mind of the child as a *tabula rasa* upon which life writes its own record was not conducive to the study of child development as a self-initiated and internally directed process. Locke's point of view dominated scientific thinking for at least two centuries, and its influence is still apparent in much of the educational literature of today. But by the latter part of the nineteenth century, the work of such men as Darwin, Galton, Spencer, and others began to make itself felt. The doctrine of "innate ideas" about which Locke had waged so gallant a fight was by this time pretty much a dead issue, but a corollary to the *tabula rasa* concept, which Locke himself had probably never intended to be drawn—that all minds are potentially alike at the start and that their later differences are entirely the result of differing experiences—had been tacitly accepted by most people. But the work of Galton and Pearson on mental inheritance and the widespread interest in Darwin's theories now rendered it essential that more precise information be gained about the early abilities and behavior of children.

Darwin himself (1877) sent the ball rolling by keeping a daily record of the behavior of his infant son, a procedure which had been followed by Tiedemann (1787) in Germany three quarters of a century earlier but which had then attracted little attention because the scientific world was not ready for it. With the newly awakened interest in the ontogenetic and phylogenetic bases of development, the possibilities of using as sources of scientific material biographical records of the development of individual infants from the time of birth became apparent. During the period extending roughly from 1885 to 1915, a fairly large number of such studies appeared in the psychological literature. Among the best known are those by Preyer (1882), Scupin (1907), Stern (1914), and Shinn (1899). Studies of the development of special forms of behavior, such as the growth of vocabulary, were also numerous at this time.

Slightly later, under the influence of G. Stanley Hall of Clark University, a second method of collecting information about child behavior and development became very popular. This was the questionnaire, a method that Hall used so extensively that it has become chiefly associated with his name, though Galton was really its originator. The obvious advantage of the questionnaire method over that of the child biography lies in its convenience, since it permits the accumulation of

large amounts of data on many subjects in a short period of time. Hall, however, had little interest in methodology for its own sake. He did not care for details; he wanted broad outlines, general principles, facts that could be applied at once to everyday life and the guidance of children. His questionnaire studies cover a wide range of topics and had a tremendous influence in stimulating interest in child development. That they violated practically every canon of questionnaire construction known to us today is of less importance than the fact that they dealt with such vitally fresh and interesting topics. Presented as they were against the background of Hall's encyclopedic knowledge of scientific facts and theories, the results of these studies attracted widespread attention and brought to Hall's laboratory for graduate study many of the most promising young men of the period. Terman, Kuhlmann, and Gesell were among the many outstanding psychologists who took their doctorates with Hall and whose subsequent work in the field of mental testing is too well known to require comment here.

A third method of child study which became prominent at this period was the study of child products, especially children's drawings. Both longitudinal studies of the development of drawing in individual children over a period of years and large-scale studies by the cross-sectional method were made. Interest in this field was accentuated by the fairly widespread belief in the theory of recapitulation, especially that aspect of it known as the "culture-epoch theory," which held that the progress of primitive man toward civilization is mirrored in the development of children from infancy to maturity. Thus it was believed that a key to the habits, customs, and ways of thinking of our early ancestors might be had from observation of the behavior and interests of present-day children.

From the standpoint of their effect on the testing movement, these early studies of child development are important for at least three reasons:

1. The clear demonstration which they afforded of the existence of behavioral sequences, reasonably uniform from child to child in pattern and order of development, which showed beyond reasonable doubt that meaningful tests of development could be devised.

2. The further demonstration that wide individual differences exist among children with respect to the age at which the various sequential stages are reached in spite of consistency in the usual order in which they occur. This called attention to the need for tests by means of which the relative position of a given child among his mates could be more precisely ascertained.

3. The specific facts ascertained in the study of these developmental

sequences suggested many useful ideas which were later incorporated into standard test items.

Thus by the dawn of the twentieth century the idea that the intellectual capacity of children and adults is subject to appraisal by uniform and objective methods had become fairly well established, even though no really useful method of making such appraisals had as yet been worked out. But the groundwork had been laid, the problem had been clearly posed, and the need for its solution was urgent. The next chapter will describe some of the early attempts at devising tests of mental ability and will show how the enthusiasm engendered when some degree of success was finally attained gave rise to overevaluation of the new methods by many persons and to the promulgation of a number of erroneous concepts which were based in the main upon artifacts arising from imperfections in the methods then available. Although more recent investigations brought to light many sources of error in these early studies, a good many of the results then obtained found their way into the scientific literature of the time and are still accepted as valid by persons who have failed to keep abreast of the more recent trends of psychological investigation.

The Early Tests (1887-1915)

INFORMAL ATTEMPTS AT APPRAISING INTELLIGENCE

No one who has much to do with children or adults can long remain unaware of the fact that they differ in ability. Consciously or unconsciously, most people form judgments about the mental capacity of the persons with whom they are associated. Many people attach a good deal of confidence to their own estimates. Yet when questioned as to the basis on which these opinions have been formed, few are able to cite any clear-cut body of evidence in support of their opinions. Some are content with reporting one or two simple episodes which to them seem sufficient proof; others can make only vague statements. Many still refer to physiognomy: "You need only to look at him to see that he is bright (or stupid)."

Different people employ different criteria for making their estimates. It thus becomes necessary to know the basis of judgment used by each person before trying to combine the estimates of different persons. Even then the procedure is hazardous. Standards that are described in similar terms often prove to be very dissimilar on closer examination. Two persons may say that they judge intelligence in terms of the ability to learn. But if one of these persons thinks of learning chiefly in terms of memorizing by rote while the other has in mind the understanding of general principles and the application of these principles to new problems, it is not likely that they will agree very closely in their ratings of individuals. Both are likely to be influenced in their judgments by such factors as the personal attractiveness of the subject, the general pleasantness or unpleasantness of their dealings with him, the social standing of his family, his physical size (large children are likely to be judged by more advanced standards than those applied to children who are small for their age, because of an unconscious tendency to regard them as older than they actually are and to compare them with children of the same size rather than with those of the same age).

Without special training few people are able to make use of any quantitative method in expressing the *degree* of a child's acceleration or retardation, even when they are able to state the *direction* in which he diverges from the average. As we saw in a previous chapter, the use of age as a rough-and-ready standard of reference was not uncommon a century ago. Probably, if the facts were known, we should find such statements as "no more sense than a baby," "he acts like a five-year-old," "he's only six but he can figure things out as well as most children of ten or twelve" in common use in everyday intercourse many centuries before we have any record of it. Such statements, to be sure, have a crudely quantitative significance, but their usefulness as measures is limited by the fact that the people who make them have, as a rule, only an extremely vague impression of what an average child is able to do at the age specified.

One of the last studies made by the brilliant Alfred Binet (1911) had to do with the ability of teachers to judge the intelligence of their pupils and the kind of evidence they employ in making their judgments. Some forty answers were received to a questionnaire sent to elementary school teachers in Paris. Two questions were asked: (1) In about what proportion of cases do you think you have erred in judging the intelligence of your students? (2) What means do you employ in making your estimates? Some thought they were never wrong; the most cautious admitted that they might be in error as often as once in three times. The methods described varied greatly from one teacher to another. Facial expression was a criterion often mentioned, as was also the shape of the head. Nearly all referred to the quality of the child's classwork. A few said they were in the habit of conducting informal interviews in which the children were asked questions of the teachers' own devising, but the questions used by the different teachers varied greatly. As a rule they were not very well chosen in any case nor was the procedure uniform for all. A varying amount of help was given as the teacher thought it was needed; that is, if the child's performance did not correspond to the teacher's previously formed opinion of his ability, an attempt was made to make it correspond. Many poorly trained "clinicians" do the same thing today!

At Binet's invitation, three teachers later came to his laboratory to demonstrate their methods to him. Each teacher spent an entire afternoon with five children whom she had not previously known. The teachers were free to use any method they wished as a means of appraising the children's intelligence. Under these circumstances, all resorted to some form of questioning, but the questions put to different children by the same teacher were frequently not the same. In a number of

instances, when the same question was asked and identical answers were given, the appraisal of these answers by the teacher differed accordingly as she had previously decided that the child in question was bright or dull.

In spite of its date, this study remains one of the best demonstrations of the many sources of error in the use of informal methods of appraising intelligence. Not that such judgments are always in error; they may be, and sometimes are, more accurate than standardized tests. But whereas the validity of a test can be appraised once and for all on the basis of its correlation with a suitable criterion or series of criteria, the validity of informal estimates will vary from person to person and even, to some extent, with the same person on different occasions. Thus one can never know how much confidence may be placed in them. Moreover, because of the lack of quantitatively expressed standards of reference, such judgments can never be truly quantitative. This lack of precision renders them less useful in the practical guidance of children and of little value for scientific research.

The emphasis which the teachers who took part in Binet's experiment placed upon bodily signs is not surprising in view of the long history of this idea. That some degree of correlation exists between physique and intellect is unquestionably true. In practically all scientific investigations a small relationship between measurements of various bodily dimensions and the results of mental tests has been found. With equal uniformity, this correlation has been too low to be of much service for individual prediction. It is most clearly seen at the lower extremes of intelligence, where the existence of many organic cases is likely to make the superficial resemblance seem very impressive. Binet himself at one time entertained the hypothesis that bodily characteristics might be found useful indicators of intellectual traits and conducted a large number of investigations to see if such signs could be found. He concluded that, on the average, a relation between mental and physical traits exists, but the rule has so many exceptions that dependence upon it would mislead almost as often as it would inform. His results are in complete agreement with the findings of most scientific workers today.

But the belief that abilities can be judged and behavior predicted on the basis of appearance is hard to eradicate. It has existed for centuries and, phoenix-like, some new theory or device appears to spring from the ashes of each old one. In part its persistence is justified on the basis of some small grain of truth. In greater part the belief may be attributed to wishful thinking. To many, physical characteristics seem more real than qualities of thought or action because they are more immediately perceived by the senses. Accordingly, these persons are more

ready to rest their confidence in such indications. Furthermore, it is flattering to the ego to feel that one has found a key that gives quick and easy access to the secrets of his neighbors' minds. By comparison, mental testing seems stodgy and laborious.

It would not be profitable at this point to review in detail the many investigations of the relation between mental and physical traits that have appeared in the literature. Paterson (1930) has performed this task very adequately as far as the quantitative studies appearing before 1930 are concerned, and a few of the more recent ones will be mentioned in a later chapter of this book. But because the earlier ideas on this subject had an important bearing on the development of mental testing, it will be worth our while to consider them in their role as a background from which the early tests were to emerge.

THE SEARCH FOR BODILY SIGNS

As early as the fourth century B.C. the possibility that the key to personality diagnosis might be found through a careful study of bodily conformation and facial characteristics was engaging the serious attention of such men as Aristotle and his followers. Although the essay on physiognomy included among the collected works of Aristotle is ascribed by most modern authorities to one of his disciples rather than to Aristotle himself, there seems to be little doubt that it was at least prepared under his influence. That the idea was not a new one, even at this early date, is evidenced by the many references to the opinions of other authorities which occur at frequent intervals throughout the essay.

Whether the *Physiognomonica* was written by Aristotle himself or by one of his disciples, it bears the unmistakable stamp of his incisive thinking. The author does not content himself with outlining a system of bodily signs by which behavioral tendencies may be recognized. Aware of the danger that circumstantial factors affecting individual experience may lead to the formation of erroneous generalizations, he asks: Are there any accepted scientific principles which might account for the supposed relationship between physiognomy and behavior? In answer he notes the following as possibilities worthy of further study:

1. Inferences based upon the resemblances of the form and features of the person in question to animals with known behavioral characteristics. Thus, a man with a deep and roaring voice, shaggy hair, powerful chest and shoulders, and slender loins might be expected to be courageous like the lion whom he resembles in body.

2. Inferences based on resemblances to the several races of mankind. If the different human races show characteristic trends of personality or

conduct, a man whose features resemble those of some race other than his own may be thought likely to show corresponding deviations in behavior.

3. Inferences derived from observation of the characteristic changes in facial expression corresponding to certain states of mind or feeling. A man whose usual expression resembles that commonly seen during states of depression or anxiety is judged to be of a melancholy disposition; the one whose habitual expression suggests mirth is thought to be of a sanguine temperament.

Each of these hypotheses, says the sage, is of doubtful validity. Many animals are noted for their courage but they differ greatly in bodily form. Moreover, few bodily characteristics are peculiar to one genus; most are common to many, "and of what use is a common attribute?" The same difficulty is found when racial characteristics are used as criteria. In respect to the judgments based upon emotional expression, it is pointed out that similar expressions may reflect very different states of mind and that different people express their feelings in diverse ways. Thus great caution is necessary in any attempt to infer mental or personal traits from bodily signs. Although the main part of the essay is devoted to an account of these signs which the author regarded as most nearly dependable, there are repeated cautions as to possible sources of error in making such inferences which militate against their practical use.

Belief in physiognomy was general throughout the Middle Ages but little of scientific importance was added to the Aristotelian outline until the close of the eighteenth century, when Franz Joseph Gall first propounded the hypothesis that personality characteristics of all kinds, including special talents and deficiencies, may be inferred by careful examination of the contour of the skull. The theory seemed reasonable since by that date the predominant role of the brain in controlling and directing behavior was generally known. Since the skull contains the brain, the inference that the characteristics of the organ might be judged by its envelope was natural enough. Although Gall's phrenological lectures, which were begun in Vienna in 1796, were prohibited in 1802 on the ground that they were dangerous to religion, the idea had already taken firm root. Gall's original skull map, which showed the location of twenty-six "organs," the development of which was presumed to correspond to the prominence in the personality make-up of as many different "traits," was later elaborated and refined by two of his students, Spurzheim and Combe, who organized these areas into two great groups—the organs of feeling and the organs of the intellect. Each of these was further subdivided into a number of subcategories made up, in their turn, of various specialized areas.

Although the completely fallacious nature of phrenology has been repeatedly shown, the idea still seems so plausible to many people that it is hard to dislodge. Scientific evidence makes slow headway against wishful thinking. Unquestionably it would be very convenient if we could appraise the abilities and analyze the behavioral tendencies of people by measuring their skulls. Unfortunately for the phrenologists, however, we now know that the contour of the skull does not conform very closely to that of the brain within it. Even if it did, our problem would not be solved, for mental functions are not carried on within small and highly localized regions of the brain but involve large areas. Incidentally, it may be mentioned that such tendencies to localization of functions within certain parts of the brain as exist do not conform to the phrenologist's skull maps in the least.

Not only the shape of the head but many other bodily characteristics have been explored in the hope of finding external signs by which personality tendencies may be recognized. During the latter part of the nineteenth century, Lombroso, an Italian anthropologist, advanced a theory that persistent criminals are characterized by certain bodily anomalies which set them off from the general population and make it possible for the skilled investigator to recognize them. Although Lombroso's theories have been generally discredited, at least as far as their diagnostic usefulness is concerned, the idea seems well founded that physical peculiarities or blemishes may play a part in the production of personal maladjustments that sometimes lead to crime. Of this and other more recent theories on the relation between mental and physical traits more will be said in a later chapter.

Although the early search for physical signs of mental characteristics was largely unproductive of direct results, it nevertheless played an important part in the early history of mental testing. The fact that its premises were wrong, its methods unsound, and the conclusions reached often absurd is of small importance in comparison with the faith that it aroused in the possibility of securing objective measures of individual potentialities in advance of actual trial. We must not make the mistake of supposing that because phrenology and physiognomy were later exploited by charlatans that they were from the beginning the work of tricksters. On the contrary, they should be regarded as examples of honest but mistaken efforts to develop diagnostic instruments at a time when the methods of psychological experimentation with which we are now familiar were scarcely known and there was little appreciation of the importance of objective verification of theories or of the nature of scientific evidence. Nevertheless, the phrenologists and their kind performed a service of great importance for the psychologists who were to come

after them. *They prepared the minds of the general public to accept the idea that mental abilities can be measured.* Probably few of us have realized what a tremendous advantage the early testing movement acquired from this preliminary breaking down of the barriers of disbelief.

EARLY EXPERIMENTS WITH TESTS OF A SINGLE KIND

SENSORIMOTOR TESTS

The "father of mental testing" was certainly Sir Francis Galton. As early as 1882 he established a laboratory in London where, for a small fee, individuals might come for a series of physical measurements including tests of sensory acuity and reaction time. These were not, of course, new measures; that which was new was the idea that a knowledge of his standing would seem interesting and important to the individual who was measured. From the time of his publication of *Hereditary genius*, Galton's interest had centered about the study of individuals rather than of groups. In 1883 he published *Inquiries into human faculty*, which is a series of separate essays, many of which had previously appeared in scientific journals. This includes his important work on mental imagery in which he shows how greatly people differ, not only in their ability to recall scenes and objects vividly and in detail but also in the devices which they employ as aids to memory. It also includes his study on the resemblances of twins. After the establishment of his anthropometric laboratory, Galton devoted a good deal of time and thought to the perfecting of instruments for increasing the precision and range of his measurements. Some of these are still in use in modern psychological laboratories.

We have already seen how Cattell, while still at Leipzig, broke away from the Wundtian tradition and turned his attention to the study of individual differences. His doctoral thesis on this topic, done under Wundt but without his blessing, was completed in 1886. Cattell's special interest at that time was in reaction time, which he, in common with Galton and many others, regarded as essentially a measure of intelligence.

Thus during the late eighties the seeds of the new psychology of individual differences were planted in soil which for many decades had been in slow preparation to receive them. But more than twenty years elapsed before the fruits of these enterprises were to ripen into useful products. Meantime, large quantities of "false fruits" appeared—false in the sense that they did not serve the purpose for which they were designed although they did increase our knowledge of the course of human development and of some of the factors by which it is affected.

After receiving his doctorate, Cattell returned to America with his interest in the measurement of individual differences still undiminished. Since he was a university professor it was natural enough that this interest should focus upon the question of the differences among students in ability to do college work. In 1890 he published in *Mind* an article that has since become a classic inasmuch as in it the term "mental tests" is employed for the first time in the psychological literature. Equally important is the fact that the article goes on to describe a series of tests which he was then actually using with his students at the University of Pennsylvania in the hope of finding a practicable method of appraising their abilities which would enable him to advise them beforehand as to their chances of college success. These tests consisted of measurements of keenness of vision and hearing, color vision, sensitivity to pain, color preferences, reaction time, tests of rote memory, mental imagery, and the like. The tests were given to each student individually and the results of each test were treated separately; there was no attempt to combine them into a single measure. The modern student will not be surprised to learn that when, some years later, the results of these tests were correlated with college grades, no significant relationships were established.¹

Both Galton and Cattell, as well as many of their contemporaries, regarded sensory and motor manifestations of the simpler kind as coextensive with the highest manifestations of which the intellect of man is capable. They regarded them as lower and higher rungs of the same ladder and believed that a dependable estimate of the latter could be had by measuring the former. This opinion was strengthened by the observed fact that idiots and imbeciles are usually slow and clumsy in their movements, are relatively insensitive to pain and blunt in their perceptive abilities.

Then as now, American psychologists were particularly attracted by the idea of testing. Psychology in America has always leaned strongly to the practical side, and, as we have seen, educational and social progress in the nineties pointed to the need for some more adequate method of appraising individual mental ability. Almost immediately after the publication of Cattell's 1890 article, a number of other psychologists decided to try out the method. Jastrow (1892) used a similar series with students at the University of Wisconsin and in 1893 set up a special exhibition, something like Galton's London laboratory, at the World's Columbian Exposition in Chicago. Visitors to the exposition were invited to come in and try their abilities. The tests were also applied to school children and the results compared with teachers' estimates of their mental acuteness.

¹ See report by Wissler (1901). The data are based upon tests given to Columbia University freshmen after Cattell took charge of the laboratory there.

Boas in 1891² seems to have been the first to do this, but the most carefully controlled tests of children were made by Gilbert in two studies which appeared in 1894 and 1897. The measures were of the same general nature as those used by Cattell. They consisted of tests of visual and auditory acuity, speed of tapping, reaction time, and several others including anthropometric measurements. The significant thing about Gilbert's work is the fact that he endeavored to introduce some control over the selection of his subjects. Each of his tests was given to approximately fifty children of each sex at every age level from six to eighteen years. Comparison with teachers' judgments was made by drawing curves showing the average score at each age for children whom their teachers judged to be dull, bright, or of average intelligence. Again, little more than a chance relationship was found.

SINGLE TESTS OF A MORE COMPLEX NATURE

While Galton and the American psychologists were wrestling unsuccessfully with the problem of inferring complex abilities from simple ones—a task which may be compared to that of inferring the nature of genius from the nature of stupidity or the qualities of water from those of the hydrogen and oxygen of which it is composed—a basically different theory and method was being developed on the Continent of Europe. Although the great leader of this movement was Alfred Binet, he was not alone in doubting that the aspects of thought and of overt behavior which are of prime importance in differentiating the genius from the idiot, or the successful man of affairs from the wielder of pick and shovel, are shown by differences in sensory acuity or speed and precision of muscular movement. Binet and his colleagues believed that differences in the ability to think and reason, to solve difficult problems by methods less cumbersome than actual trial and error, to make use of the experiences of the past in adapting to new conditions can best be measured by setting the subject problems that involve these very processes. In a word, they were in favor of approaching the whole question of mental measurement by means of a sampling technique. The immediate question which they had to solve was accordingly that of finding suitable tasks which could be fairly regarded as samples of the kind of abstract judgment and reasoning demanded in those situations which observation had shown could be handled by “intelligent” persons but not by the “unintelligent.” There was not, at that time, very much in the way of clearly formulated opinion as to the nature of these situations; as a matter of fact, the

² Boas did not publish the results of this experiment himself, but in 1892 they were analyzed and the findings presented by T. L. Bolton in the *American Journal of Psychology*.

development of testing devices and the definition of the characteristics to be measured have always gone hand in hand. But then, as now, most people had some general concept of the nature of the differences that divide the bright from the dull.

But the influence of faculty psychology was still too strong to admit the idea of devising a single scale in which samples of many different aspects of mental ability would be combined in order to provide a rough but serviceable means of appraising "general" intelligence. Even those psychologists who rejected the idea that such forms of mental activity as solving mathematical or mechanical problems, forming abstract judgments, or comprehending abstract relationships are merely more advanced forms of the simpler processes studied by Galton and Cattell, were not yet prepared to give serious consideration to the idea of a single measure of intelligence. On the other hand, the use of a number of different tests with the same children in order to see how closely the results of one compared with those of the others and with other aspects of ability was fairly early in appearing.³ In 1891 Münsterberg proposed a long series of tasks which he stated he had used with school children. The following are examples: giving the colors of ten things from a printed list of their names as "grass—green"; naming colors; counting angles in irregularly shaped polygons; single-column addition; and judging the length of a line in terms of multiples of a given line. All these tests were scored on the basis of speed and number of errors; a method still used in many of our nonverbal tests.

In 1896 Binet and Henri published an article in which they described tests which they proposed to try out with school children and which were designed to "measure" each of eleven named "faculties" or mental processes: (1) memory, (2) mental imagery, (3) imagination, (4) attention, (5) comprehension, (6) suggestibility, (7) aesthetic appreciation, (8) force of will as indicated by sustained effort in muscular tasks, (9) moral sentiments, (10) motor skill, and (11) judgment of visual space. For each of these, a number of different tests were proposed in order to cover different aspects of the ability to be measured. This article was the first of a long series in which these and a number of other devices were painstakingly tried out with school children of different ages to ascertain, first, the extent to which the scores improved with age and school attainment; and second, whether or not there were any consistent differences in the performance of children whom their teachers regarded as bright or dull. Some of the measures stood up well under these trials;

³ Except for the fact that no way of combining the results from the entire series into a single measure had been devised, the method then used is not unlike that employed in most of the modern group tests of intelligence.

others showed so little differentiation between the groups that it was thought not worth while to experiment with them further. In all cases, not only quantitative differences, such as the ability to repeat from memory a longer series of digits or to tap more rapidly with a stylus, but differences in the type or manner of the children's responses were noted. Here we see a highly important difference between the method followed by Binet and that used by most of his contemporaries. The latter, for the most part, experimented with tasks in which the responses could be graded in terms of such physical units as time, space, or number. Such measures as the average extent of the errors made in judging the length of lines, the average number of taps made in a given period of time, or the strength of grip as measured by the number of pounds of pressure registered on a dynamometer were preferred because the results could be expressed in directly quantitative terms which permitted easy comparison of age groups or of groups selected on any other objective basis. But as Binet's experiments progressed, we find him discarding more and more of these "objective" tasks in favor of those measured in an "all-or-none" fashion. The validity of the latter as measures of intelligence was determined not in terms of an increase in a quantitatively graded measure such as time or number but in terms of *an increase with age, school grade, or estimated intelligence in the percentage of children who solved a given task in a certain way*. It is important to note that the kind of solution which received credit, according to Binet's system of scoring, was not always that which the average adult would consider the best one, although in many instances this proved to be the case. But Binet was never content to use armchair judgment as a criterion. Always, he turned to the actual performances of children for his evidence. Thus he noted that when asked to tell what they can about pictures, young children simply enumerate objects while older ones tell what action is depicted. He observed that very young children are unable to define familiar nouns in any terms at all, even though they use them freely in everyday speech, while by the age of five or six, children typically give definitions in terms of use and some years later in terms of genus. He found that it is much more difficult to copy a diamond than it is to copy a square, even though they have the same number of sides and angles. All these and many other findings tended to strengthen the view which Binet had already put forth in an article published in the *Revue philosophique* (1898), in which he emphasized the fundamental differences between the application of measures of material objects as used in the physical sciences and the application of such units to psychological processes. He noted that a zero score on some test which its originator has designed to be a test of intelligence does not mean that the subject has zero

intelligence. He called attention to the findings of Weber and Fechner as evidence that changes in the physical magnitude of a stimulus-object are not accompanied by changes of equal magnitude in the sensations aroused. Mental measurement must therefore be expressed in psychological rather than in physical units.

Binet's interest in what we commonly speak of as the "higher" or "more complex" mental processes was shared by many others. Ebbinghaus's classical study of memory had appeared even earlier (1885) than the first of Binet's publications in this field. Ebbinghaus, to be sure, was not concerned with measurement for its own sake nor with the study of intelligence in a broad sense. He wished to learn something about the factors that influence memorizing and forgetting. The problem which he set for himself was that of drawing up a series of general principles by which memorizing in its purest form, that is, the form in which it is least affected by such circumstantial factors as contrast, meaning, or specialized interests and experiences, takes place. To do this it was necessary for him to devise a suitable method. Eventually he decided upon the use of nonsense syllables, a device which is still popular in the psychological laboratories of today. Lists of nonsense syllables to be memorized were included in some of the early series of mental tests but they are rarely used for this purpose at present.

The success of Ebbinghaus's method led quickly to the devising of other "memory tests." These included tests for measuring the span of visual apprehension, memory for digits, for lists of words, for sentences, and so on. In 1900 appeared Stern's *Über die Psychologie der individuellen Differenzen*, in which a number of tests for appraising both simple and complex abilities were described.

Thus by the turn of the century, considerable progress in the devising of various kinds of tests and test items had been made. But with increased awareness of the wide range of differences in *pattern* of ability that exist from child to child, the need for some more convenient form of expressing the general mental level of the individual became apparent. What was wanted was something in the nature of a synopsis—a short series of tasks that would take account of the varying patterns of ability possessed by different persons. Such a test would not give an unfair advantage to those whose abilities run along one line while handicapping those with a different alignment of talents. Its range of content, however, should not be so great as to make for a superficial type of appraisal. No single test of reasonable length can possibly measure all the characteristics of any human being.

Even before the publication of Binet's 1905 scale, a number of persons had been groping toward the concept of such a testing device though

without clear formulation of their aims. Generally, the procedure consisted of substituting a series of questions, often not very well selected but covering a wide range of topics, for the more precisely formulated tests of single aspects of ability that had formerly been used. In some ways this seemed like a backward step, for it certainly meant a considerable loss in accuracy of measurement. Nevertheless it represented a vague groping toward the concept which the modern statistical worker would formulate by stating that "validity" must take precedence over "reliability." Most of the tests in common use had shown only low correlation with such indications of ability as school success or teachers' judgments of intelligence. As a result, many of those who had at first been most enthusiastic about the possibilities of mental testing as a practical device for the guidance of children and college students lost interest in the subject. In America, particularly, the interest that had flamed so high in the nineties was reduced to a feeble and intermittent glow during the next decade. Binet, on the contrary, never lost faith. By the beginning of the new century he had accumulated a tremendous amount of data about the way children respond to a great number of different kinds of tasks. He had compared the responses of individual children with the quality of their schoolwork and with such other evidences of their abilities as he was able to secure. He knew which of his measures looked promising and which appeared to be worthless for his purpose. He had formed a pretty clear idea of what children of each age are able to do. He thus had the raw material for constructing a scale, but no one, thus far, had devised a plan for putting such material together. Binet had to be his own architect.

THE 1905 SCALE

From the beginning of his professional life, Binet's strong humanitarian leanings had given him an interest in educational problems, particularly those having to do with the education of retarded children. He was an early advocate of special classes for those unable to profit by the instruction given in the regular grades. When, in 1904, the Minister of Public Instruction in Paris appointed a commission to decide what measures could be taken for the education of such children, it was thus but natural that Binet should be consulted. It was decided that a special school or schools should be established and that all children who, on the basis of medical and pedagogical examination, were found to be incapable of learning by the ordinary methods should be removed from their regular classes and given special instruction in accordance with their needs. Evidently a means of making the selection was called for, and it was specifically to meet this need that Binet and his colleague, Th. Simon,



FIG. 2. ALFRED BINET. (Courtesy of the C. H. Stoelting Company.)

constructed their first formal scale for appraising the intelligence of children. This scale differed in a number of ways from most of those which had previously been constructed. The most important differences were these: (1) It made no pretense of measuring any single type of ability or "faculty" in a precise and adequate manner, but aimed instead at getting a general idea of the child's mental development along as many different lines as possible through setting him a wide variety of tasks. Binet and Simon specifically pointed out that such a procedure is not measurement but merely a device for classification. (2) It required only a short time for its administration. Binet was a practical man who had had much direct experience with children. He recognized that fatigue may affect a child's performance and he also knew the practical difficulties arising from an attempt to examine a large number of children by means of a long and time-consuming test. (3) Unlike their predecessors, Binet and Simon selected their test items with a reasonably clear and definitely molar idea of the nature of intelligence. They pointed out that any attempt to measure all the sensory, motor, perceptual, and other elements that are involved in what is commonly known as an intelligent act would be prohibitively time consuming and probably unnecessary. Rather, they attempted to move directly toward what they regarded as the essential factor of intelligence—the ability to make sound judgments. "To judge well, to comprehend well, to reason well, these are the essentials of intelligence. A person may be a moron or an imbecile if he lacks judgment, but with good judgment he could not be either," they affirmed. (4) Instead of arranging their tests according to general type, that is, putting memory tests in one group, numerical problems in another group, and so on, they arranged the tasks in order of difficulty without regard to their apparent similarity or dissimilarity. The order was approximate only, since pressure of time demanded that the scale be ready for use as soon as possible. Such standardization as was done was based upon the performances of fifty normal children ranging in age from three to eleven years and a number of subnormal children from the Salpêtrière as well as a few retarded children from the primary schools.

The scale included thirty items, ranging in difficulty from those suitable only for the classification of very low-grade idiots to those intended for children in the upper elementary grades. The following examples (in which the instructions for giving and scoring have been omitted) will give a general idea of the kind of tasks which, at this time, Binet thought indicative of various levels of intellectual ability. The numbers are those of the original scale. They represent increasing difficulty, but needless to say were never regarded as having more than serial value.

1. Coordination of movements of head and eyes in following a lighted match.

6. Execution of simple orders and imitation of gestures.

(These are the easiest and most difficult of the series found to differentiate the idiots from the imbeciles.)

9. Enumerating objects shown in pictures.

15. Repetition of a sentence of fifteen words after a single hearing.

(Test 15 marks the upper limit of imbeciles.)

20. Stating similarities of two familiar objects.

25. Supplying the missing words in easy sentences (a device originally used by Ebbinghaus).

30. Distinguishing between abstract terms.

In using these tests for the selection of candidates for special classes, two practices that are worthy of special note were followed in choosing the children to be tested. First, instead of merely asking teachers to name children whom they thought should be placed in such classes—a plan which, had it been followed, would inevitably have resulted in the selection of troublesome children rather than retarded ones—teachers were asked to prepare lists of all the children in their classes together with their ages and school marks. Overage children with low marks were then chosen as probable candidates. But in order that the children might not guess that the dullards were being chosen for testing, and also in order that the examiners might not be influenced in their procedures by a knowledge of the child's school retardation, teachers were asked to distribute a number of entirely normal children among those sent for testing. This practice guarded against the "halo effect" about which we hear so much today but, alas, so rarely take practical measures to avoid. It also provided data which were found extremely useful in revising the scale later on.

THE 1908 SCALE

In 1908 Binet and Simon published a new scale, based in part upon the tests used in the 1905 scale but with the addition of a number of new items and the elimination of others which had proved to be of less value. The outstanding feature of this scale, however, lay in the fact that instead of being arranged in order of difficulty, the tests were now grouped according to the age at which they were commonly passed. A new term, which was to become one of the most useful and familiar in the entire literature of mental testing, was here introduced for the first time—

"mental age." A child's test standing was no longer to be stated simply in terms of the number of items passed but in reference to age standards. Just as the statement that Peter takes an eight-year size in suits or weighs as much as an average ten-year-old gives us a more exact idea of his size than does a mere statement of his chronological age with no indication as to whether or not he is of average size, so a statement of a child's "mental age" tells us more about his level of mental development than does merely a statement of his chronological age.⁴ The method of finding the mental age by means of the 1908 scale was first to assign to the child a mental age corresponding to the age level at which not more than one test was failed. To this basal age was to be added one year for each five tests passed at levels above the basal. Inasmuch as the number of tests at successive ages varied from three to eight and as no credit for fractional years was allowed, it is evident that the method could at best yield only very crude results. But however imperfectly it was worked out at this time, the idea was a notable one destined to change the course of mental testing for many years to come. A second major difference between the 1905 and the 1908 scales was the omission from the latter of the tests designed for use with idiots. There were two reasons for this. Chiefly there was the fact that his work in the schools brought about a gradual shift in Binet's interest from that of diagnosing subnormality to obtaining a better understanding of normal children. As was pointed out in the last chapter, French psychology at this time was largely concerned with the abnormal. It was therefore natural enough that Binet's interests should follow a similar pattern. By 1908, however, he was beginning to see that when applied to the rank and file, his tests might have a far wider field of usefulness, for if teachers could be brought to see that nine-year-old Sam whose mental age was but seven should be treated as a seven-year-old, many pedagogical difficulties would disappear. A second reason for the elimination of the "idiot scale" was Binet's realization that such tests were in most cases superfluous, for the idiot can usually be recognized as such without the aid of formal testing.

It should perhaps be noted that Binet was not the first to devise a scale in which tests were grouped according to the usual age of passing them, nor was he the first to show the advantages of age as a standard of comparison. In 1887, twenty-one years before the appearance of the 1908 scale, Dr. S. E. Chaille published in the *New Orleans Medical and Surgical Journal* a series of tests for infants up to three years of age

⁴ Provided, of course, that a suitable test is used and that it is properly administered and scored. The inaccuracies in Binet's 1908 scale make it questionable whether this statement would hold good for the results of that scale, except for children much accelerated or retarded.

arranged according to the age at which they are commonly passed. Although Chaille does not use the term "mental age," the idea is clearly implied in his discussion of the manner in which the tests were to be applied and interpreted. Because the journal in which the paper appeared had, for the most part, only a local circulation and also, perhaps, because the scientific world was not yet ready for it, the significance of this early study was not realized until half a century had elapsed since its publication.

THE 1911 SCALE

Shortly before Binet's untimely death, a second revision of the Binet-Simon scale appeared. The 1911 scale differed from that published in 1908 in its details rather than in its major principles. With the exception of Year IV, five tests were provided at each age from three to ten years, with additional groups of five tests each for ages twelve and fifteen years and for the adult level. The method of scoring was the same as that used in the 1908 scale except that the additional allowance for tests passed above the basal year was 0.2 year for each test passed, thus permitting the use of fractional parts of a year in computing the mental age. Binet himself seems to have been reluctant to recommend this practice, and expressed himself as being very doubtful whether the instrument was sufficiently delicate to warrant such a refinement in its use.

Other changes from the 1908 scale involve shifts in the location of certain items that had been found to be incorrectly placed, elimination of some items and the addition of others thought to be more diagnostic, and greater precision in the instructions for administration and scoring.

THE USE OF THE BINET-SIMON TESTS IN AMERICA

*GODDARD'S TRANSLATIONS*⁵

In 1906 the Training School at Vineland, New Jersey, established a special department for research on methods of studying and educating feeble-minded children and invited Dr. Henry H. Goddard to become its first director. Dr. Goddard soon realized that a major requirement for such work was a diagnostic instrument, capable of distinguishing between

⁵ Although often referred to in the literature as the "Goddard Revisions" of the Binet-Simon tests, Goddard's two scales are translations of the 1908 and the 1911 scales without real change. American coins replace the French coins used by Binet in the test of naming coins, English sentences of similar length and difficulty are substituted for the French sentences in the tests of memory for sentences. Minor alterations such as these are the only changes made.

those children who were to be regarded as normal and those of subnormal mentality, and between the different levels of subnormality found among institutionalized children. In his search for such a device he came upon Binet's report of his 1908 scale and was immediately impressed with its potentialities. With characteristic energy he set to work at once to translate the scale into English with such minor changes as were necessary to make it applicable for use with American children. After trying out the translation on enough cases to satisfy himself as to its usefulness, he published it (Goddard, 1910); in the same year he also presented a report of its application to some 400 inmates of the Vineland institution. As a check, he and his assistants also administered the test to approximately 2000 children in the Vineland public schools (Goddard, 1911). The results of the second survey not only demonstrated how great was the difference between normal and feeble-minded children in performance on this scale but also showed that the public schools were carrying an unsuspectedly great load of mentally defective children. The intellectual backwardness of these children was manifested not only in their low test scores but by their inability to do the work of their school grades. Except for the youngest ones, who had not yet had time to pile up a record of school failure, these children were in most cases greatly retarded in grade placement. With continued failure, many of them had become disciplinary problems. They were absorbing far more than their rightful share of the time and attention of their teachers and in many instances their influence upon the younger children in their classes was definitely undesirable.

This experience convinced Goddard that the new scales had a far wider field of usefulness than he had dared to hope. How much wasted effort might be saved, how much easier and more effective would be the work of public school teachers, if those children who could not be taught the ordinary subjects of the school curriculum but needed an educational program especially adapted to their limited abilities could be removed from the regular classes and placed in special rooms under the charge of teachers who understood both their handicaps and their needs! True, such classes had been tried in a number of places, as a rule with moderate—but only moderate—success. But, argued Goddard, a part of the difficulty almost certainly could be traced to the lack of a uniform and objective system for selecting cases for such classes. As a result they had tended to become filled with delinquents, with children suffering from physical handicaps, with children unable to speak English, and with all sorts of other exceptional cases. The classes were usually too few and consequently too large for one teacher to handle effectively. There was no very clear idea as to how, if at all, the training of these children

should differ from that given in the ordinary schools. Finally, little or no provision had been made for specialized training of the teachers.

All this, said Goddard, should now be changed, since the tests provided the first and most essential means for making such a change possible. Like the apostles of old, he proclaimed his belief in all possible quarters. He gave public lectures, demonstrated the method of testing before educational associations, groups of social workers, medical men. His translation of the 1908 scale was followed by a translation of the 1912 scale in the same year in which the latter appeared in France. This soon became the standard testing instrument which was used everywhere until the Stanford Revision appeared in 1916.

The need for providing some kind of special training for teachers of backward and feeble-minded children was noted by the Vineland staff at a comparatively early date. In 1903, a special summer course for their own teachers had been organized and the following year was opened to others who might be interested. Teachers from other institutions made up most of the enrollment at first, but after the publication of Goddard's translations of the Binet scale, and with the rise of interest in special classes, public school teachers and others who were particularly interested in learning to use this new instrument for classifying children flocked to the summer school. Arrangements were made for them to live in the institution. This gave them intimate daily contact with several hundred feeble-minded children of all types and enabled them to see at firsthand how the children were handled in the schoolroom and workshops. Of much greater importance, however, were the attitudes with which, almost without exception, those who attended the Vineland summer sessions at that date became fired. Tests and testing were still new. Their possibilities had been shown in an amazingly dramatic and clear-cut fashion; realization of their limitations was yet to come. The members of the teaching staff at Vineland were young, vigorous, and ardent in their belief that a key had been forged which, if properly used, might open the way to a solution of many of the social ills which had plagued humanity for untold centuries. For these were the days of uncritical application of the newly discovered Mendelian principles of heredity to mental inheritance in man. These were the days when the Kallikak family was discovered.

MENTAL TESTS AND HEREDITY

In 1900, Mendel's now classic paper, which had originally appeared in 1865, was rediscovered by De Vries and his associates. During the next decade, Bateson and others were able to show that at least in their

broad outlines, the Mendelian principles hold good for a wide variety of characteristics of plants and animals. Again, Goddard's brilliant imagination leaped ahead. If the Mendelian laws could be applied to such matters as the coat color of rabbits, the height of the garden pea, and the length of an insect's wing, why should they not also apply to the intelligence of man? Up to now the lack of a useful and objective measure of intelligence had been a major stumbling block in the way of discovering what scientific principles underlie mental variations. Now that this difficulty seemed to have been largely overcome, the time appeared ripe to lay the foundations for a truly scientific study of the origins of individual differences in mental traits. Galton had shown that genius tends to run in families but he had made no attempt to indicate the manner of its transmission. The large number of brothers and sisters and their remote relatives in the institution at Vineland had already convinced Goddard that mental deficiency as well as genius follows the blood line. Now his problem was to see if it conforms to the newly discovered principles of Mendel.

Goddard's subsequent work on this topic is too well known to require detailed comment here. His two books dealing with the inheritance of mental defect, *The Kallikak family* (1912) and *Feeble-mindedness; its causes and consequences* (1914), were widely read and discussed by psychologists, educators, and the general public. That the evidence upon which the data were based was flimsy when judged by modern standards, few people were then prepared to note, for the era of scientific criticism of tests and of testing standards had not yet arrived, nor had the study of the laws of inheritance advanced very far beyond the level at which it was left by Mendel. That Goddard's crusading zeal often blinded him to sources of error in his work, which, in the cold light of modern science, seem glaringly apparent, is unquestionably true. The assistants whom he employed to secure his genealogical records had relatively little training but were fired with Goddard's enthusiasm. That they may sometimes have tended to find mental defect where mental defect was to be expected was perhaps inevitable under the circumstances, but no one can doubt the sincerity of their attempts to get at the facts.

The importance of Goddard's work on the inheritance of mental defect does not depend merely upon the soundness of his methods or upon the factual accuracy of the reported data for his individual cases. Whether or not the "nameless"⁶ feeble-minded girl was the starting point

⁶ In a recent note, Goddard (1942) states that the girl's name was known and that the appellation "the nameless one" was merely a pseudonym used to preserve anonymity.

The Kallikak family is the report of the family tree of Deborah Kallikak, an

of the long line of defectives in the "feeble-minded" branch of the Kallikak family need not now concern us greatly. Whether or not all the cases classified by Goddard as "feeble-minded" were actually so is not of great consequence. What matters is that the book afforded the first clear demonstration of a truth that few will now question—that mental deficiency tends to run in families. Moreover, the clear-cut and dramatic style in which the data were presented, the descriptions and photographs of individual cases with their sordid records of poverty, drunkenness, and crime, made a tremendous impression upon the reader. The effect of the book upon social and educational practices and upon the attitude of psychologists toward tests and testing was profound and far reaching.

One reason why the report of *The Kallikak family* aroused such widespread interest is the stress Goddard laid upon mental deficiency as the soil from which such social ills as delinquency and crime, sexual promiscuity with its accompanying evils of venereal disease and illegitimacy, drunkenness, and pauperism, commonly spring. His figures on the relative frequency of such conditions in one of the two contrasted branches of the family were startling enough to arrest the attention of all who read them. Up to that time, few people had thought of mental deficiency as reaching into the lives of the normal population and affecting its well-being. But if, as Goddard's data appeared to indicate, the feeble-minded person is not merely a burden upon society but is likely to become an active menace as well, then it was high time that something be done about it.

What could be done? Evidently the success of any kind of preventive or remedial measure would be dependent upon the identification of the feeble-minded while they were still children, before they had had time to become hardened in the ways of the criminal. This identification, so Goddard urged, should be made on an objective basis and not be left to the fallible and biased judgments of untrained persons. It should be

inmate of the institution at Vineland. Impressed by the fact that the girl's name (*Kallikak* is, of course, a pseudonym) was the same as that of many well-known persons in the state, but was also shared by paupers and criminals, Goddard set out to ascertain whether or not there was any connection between the superior and inferior bearers of the family name. The ancestry was traced back to the time of the American Revolution, when a young soldier of good family, Martin Kallikak, was said to have had a casual intimacy with a feeble-minded girl in a tavern. From this union a child was born to whom the mother gave the name of its father, Martin Kallikak. Later, Martin Kallikak, Sr., married a normal woman and thus founded the normal branch of the family. From the illegitimate first-born son, Martin Kallikak, Jr., who, like his mother, was said to have been feeble-minded, a total of 480 descendants was traced of whom 143 are stated to have been certainly feeble-minded, 46 were classed as normal, while the mental status of the others was either unknown or doubtful. From the legitimate marriage, 496 descendants were located. There were no feeble-minded among them, no illegitimate children, no criminals.

made on the basis of mental tests. His view was shared by most psychologists, and although opposition to the tests was common among other groups, such opposition could not long hold its ground against the repeated demonstrations that the tests not only worked but provided a numerical basis for indicating the degree of the child's backwardness. To most people such a statement carried far more conviction than a mere general assertion that a given child was feeble-minded. Figures have a certain scientific aura that words lack.

An important part of the summer school curriculum at Vineland consisted in training the students to give Binet tests. Some time was also spent on the various kinds of nonverbal tests which were being tentatively experimented with in a number of places. Courses for the training of teachers of the feeble-minded as well as special courses in mental testing were soon organized in a number of universities throughout the United States. Reports of new tests, of special groups of cases tested, of the frequency of the "feeble-minded" in various populations appeared in increasing numbers in the psychological and educational journals. In 1912, New Jersey passed a law requiring that all municipalities having as many as fifteen children who were three or more years retarded in school should provide special classes for their instruction. The law also provided that the teachers of these classes should have had at least six weeks of special training for their work. The details of this training were not specified, but by common consent the institutions which set up such courses included mental testing as an essential part of the work. Thus within a very few years the number of people who had received a few weeks' training in testing, but who, in many cases, had little or no other psychological background, became fairly large. With the uncritical enthusiasm born of ignorance, they set about the examination and classification of school children. The IQ was not yet born; children were said to "test" so many years "above age" or "below age." Any child testing as much as three years "below age" was said to be "feeble-minded," a candidate for a special class or for an institution if arrangements could be made for his admission.

In *The Kallikak family* Goddard had estimated the proportion of the feeble-minded among the delinquent and criminal groups at not less than 25 per cent and possibly even as high as 50 per cent. These estimates were based in part upon the actual testing of reformatory inmates and in part upon the opinions of the superintendents of various penal institutions. The publication of these figures led immediately to further studies of the relation between mental defect and crime. Uniformly, tests either of juvenile delinquents or of adult criminals showed amazingly high

numbers, often as many as 75 per cent, who tested three or more years "below age." This discovery was both disconcerting and reassuring, for while it suggested that grave injustice had been done to a large number of persons who should not have been held responsible for their acts, it also gave reason for hope that if the problem of mental deficiency could be solved, crime would be greatly reduced.

But as more data on test results became available, questions as to the infallibility of the tests became more frequent. Children were retested, and it was found that the results of the second test sometimes varied from the first by disturbingly large amounts. It was also noted that there was a marked tendency toward a downward shift in test standing with advancing age, a fact which suggested that the standards at the older levels were too difficult. If so, since few of the cases tested in surveys of delinquents and criminals were below the age of thirteen or fourteen years, their typically low standing might be in part an artifact of the test used. Comparative studies of nondelinquents of similar age showed this to be the case. More low-testing individuals among the delinquents than among the nondelinquents were still found, but the difference between the two groups was much reduced when the factor of age was controlled. Other sources of error were noted. The procedures followed by different examiners were not always uniform, and as more people began to use the tests, often with no other guide than that afforded by the printed instructions, such discrepancies became more frequent and more glaring. Doubt began to be felt as to the significance of some of the test items. Most frequent, perhaps, of all the criticisms was the statement that the tests were too linguistic in character. Children from foreign homes with only a limited knowledge of English were, it was felt, unjustly handicapped by these tests. Another kind of scale was needed for such cases.

Fortunately this era of criticism did not set in until the usefulness of the tests had been thoroughly demonstrated. The crusading spirit that marked the adoption of the Binet tests in the United States in the years between 1910 and 1916 led to many errors of procedure and interpretation, but without its impelling force, progress in the field of mental testing would unquestionably have been much slower, and there is no guarantee that the methods adopted would have been more sound.

As we look back over the road along which we have come it is apparent at once that improvements in testing procedures have been made for the most part on an empirical rather than on a theoretical basis. Generally speaking, we have seen our mistakes more often than we have foreseen them. But progress on an empirical basis demands the

accumulation of empirical data, and as reports of tests and testing began to occupy an ever-larger space in the psychological literature, they were scanned with eager attention by psychologists, who were fascinated with the possibilities of the new instrument but were becoming increasingly aware of its imperfections.

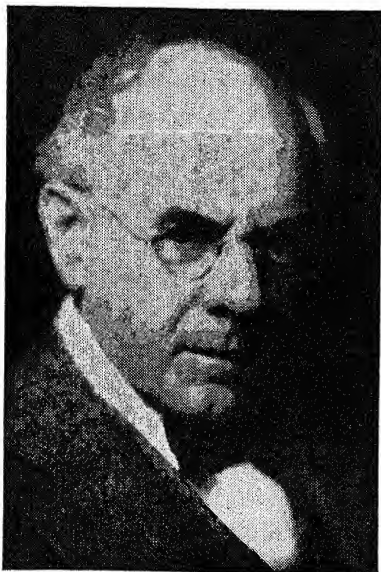
Later Developments

EARLY ATTEMPTS AT REVISING

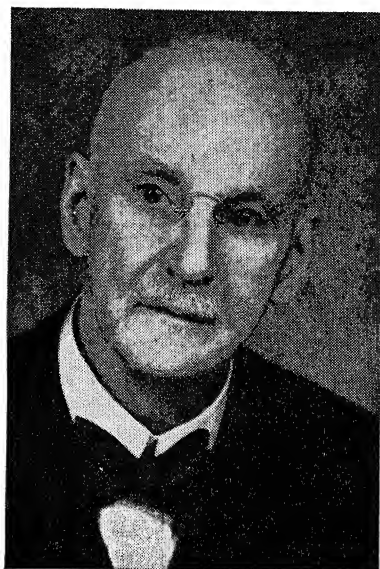
THE BINET-SIMON SCALE

Goddard was not the only psychologist awake to the importance of Binet's work, though he was the first to translate the tests into English and put them to actual use with American children. Others, both in the United States and abroad, were following the development of the new idea with interest. When, in 1911, they learned of Binet's death, the need for carrying on the work which he had left unfinished immediately became apparent. In 1911, only a few months after the appearance of Goddard's translation, Kuhlmann brought out another which differed from that by Goddard in some of its details; a year later he published a second and slightly revised form of this translation. Bobertag published a German translation in 1911, and a number of other translations and adaptations appeared at a relatively early date.¹ Of these only two need concern us here. The first is the report by Terman and Childs (1912) of their preliminary work on the Stanford (1916) Revision. The second is the Point Scale by Yerkes, Bridges, and Hardwick (1915). The latter is important because it marks a cleavage that still exists in respect to theories of test construction. Binet, it will be recalled, constructed his first scale on a point system in which the items were arranged in order of difficulty and the child's score was the number of items passed. He later rejected this method in favor of a year scale, an arrangement preferred by most of his immediate successors. Yerkes and his associates, however, pointed out certain questionable features of this method. First there was the matter of the "basal year." In a test of the year-scale type, the procedure commonly followed consists in adding to the highest year level in which the subject is able to pass all the tests, a proportional

¹ For a discussion of these early revisions, see Peterson, *Early conceptions and tests of intelligence*.

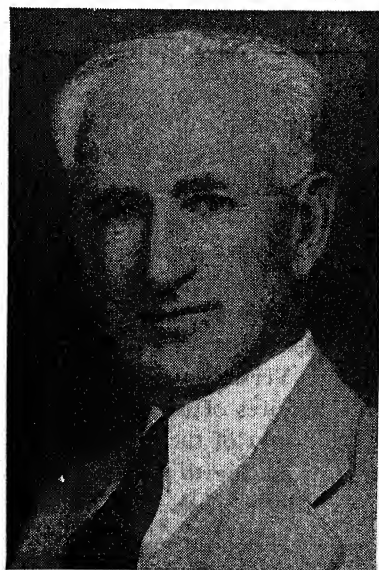


J. McKeen Cattell



Henry H. Goddard

Robert M. Yerkes



Edward Lee Thorndike

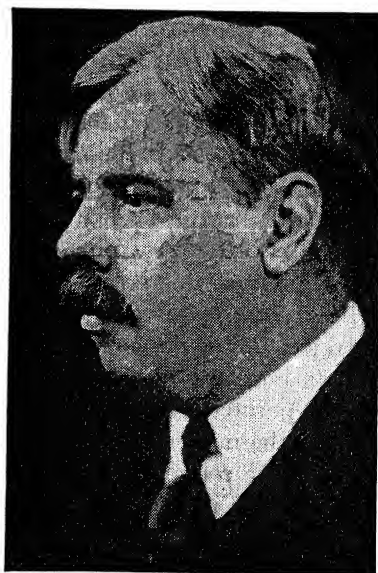


FIG. 3. SOME AMERICAN PIONEERS IN MENTAL TESTING.

number of months for each test item passed at higher age levels.² It is assumed that since the tests below the basal age are presumably easier than those placed at or above this level, the child may be given credit for them without actual trial. This practice, so Yerkes claimed, assumes a far greater uniformity in the *pattern* of intelligence from child to child than actually exists. The fact that the majority find one of two specified tasks easier than the other does not necessarily mean that it will be easier for all. The correlations between test items, so Yerkes stated, would need to be considerably higher than they are commonly found to be in order to justify the use of a method which, in effect, accepts a child's performance on one set of items as a valid indication of his probable performance on another set, especially when the two may be quite different in type. A similar opportunity for error in the appraisal of a child's ability was to be found in the practice of continuing the testing only to a point at which all the items³ were failed. A fairly large number of children, so Yerkes stated, were found to go beyond the limits set by the basal and final years in one or both directions when the limits of testing were extended.

Yerkes was also concerned over the question of marginal successes and failures. As left by Binet, many of the individual test items were made up of several parts of which the child need pass only a given number in order to receive credit for the entire item. If, for example, a given item included three questions of which two must be answered correctly, the child who passed all three received no additional credit for his superior performance, while the child who passed but one received no credit at all. By this practice, so Yerkes believed, much of the discriminative value of the scale was unnecessarily discarded. Moreover, the two sources of possible error might—and not infrequently did—tend to augment each other. Failure to allow any credit for passing less than the required part of an item at the upper extreme would mean, according to the accepted practice, that unless a child had passed other items at this level he would not be given a chance at those further on. Yet his partial success, even though insufficient to permit giving him credit for the total item, certainly would indicate an exceedingly small difference between his ability and that of a second child who passed the same item by the narrowest possible margin and would therefore be allowed to try the next series of tests. The same general principle would operate at the other extreme.

² The rule prescribed in the Goddard translation was to use as the basal the year at which all the tests except one were passed. This practice, of course, increased the error of measurement by an appreciable amount, particularly when the number of misplaced items in the early scales is taken into account.

³ Or all except one. See the preceding footnote.

Yerkes admitted the practical usefulness of the mental-age concept but pointed out that this need not be lost in a point method of scoring. All that was needed would be a table of standards showing the average number of test items passed by children at each successive chronological age. By reference to such a table, point scores could be immediately converted into mental ages. These arguments were convincing to many people, especially those with experience in testing who had had occasion to note many instances in which the errors in question actually resulted in rather gross misclassification of individual children. That the importance of these factors was to a great extent traceable to the particular scales then in use was not known. Subsequent work has demonstrated that an accurately constructed year scale will yield results that differ but little from those obtained by the use of the same items arranged and administered according to the point-scale method. But the Point Scale devised by Yerkes and his colleagues was actually a better scale than that put out by Goddard, and in spite of the prestige already gained by the latter during the four years in which it had been used, the Point Scale became its only serious rival in the United States. Had the latter not been so quickly overshadowed by the Stanford 1916 Revision, it is probable that it would have attained wider usage than it did.

THE STANFORD 1916 REVISION

In 1916, L. M. Terman of Stanford University published the famous Stanford Revision of the Binet-Simon tests. In a sense this was the first real revision that had appeared since the time of Binet. The earlier forms were essentially translations with few new features. But the Stanford Revision introduced many new items and changed the location or the method of administering or scoring of so many others that it became essentially a new scale, bearing only a surface resemblance to those that had preceded it.

The widespread popularity of the Stanford Revision, which quickly made it the standard for all testing in the United States and enabled it to maintain this position of unquestioned prestige for more than two decades, can be traced to a number of factors. Chief among these was its introduction of the intelligence quotient as the basic way of expressing test results. For some years, mental testers had recognized the inadequacy of the mental age alone as an indication of a child's position with reference to his mates, especially when the question of mental deficiency was involved. The current practice of regarding any child who tested three or more years "below age" as a candidate for a special class or perhaps for an institution obviously meant the use of unequal

standards for older and younger children. Goddard and others proposed that two years be taken as the standard for children below the age of nine years as a rough way of equalizing the disparity, but since a rigid interpretation of this rule might mean that a child who had been classed as "feeble-minded" up to the age of nine would become "normal" on reaching his ninth birthday, the idea was not generally accepted. As early as 1912, both Stern and Kuhlmann had suggested that the ratio of the mental age to the chronological age might be a useful way of indicating the extent of a child's acceleration or retardation, but as neither of them incorporated the idea as a basic feature of a mental test, the device attracted little attention till 1916, when Terman made it popular. The simplicity of the measure, the ease with which it could be computed, together with the fact that it supplemented but did not displace the mental-age concept which had proved so valuable as an aid to understanding the mental capacities of children, appealed at once to those who had been groping for a quantitative measure that would have equivalent meaning for all. Moreover, Terman did not content himself with a mere description of the new measure. He presented figures, showing what proportion of public school children had been found at each of the successive IQ levels. He presented a table which is still quoted without change in many textbooks, stating in descriptive terms how the IQ should be interpreted. Terman took care to point out that the limits set in this table were arbitrary and that the table itself was intended only to serve as a general guide to the use of the new measure, not as a series of standards that could be applied by rule of thumb. Unfortunately these cautions were soon overlooked or forgotten by many enthusiasts who were dazzled by the numerous reports of the marked contrasts in school achievement and general behavior of children with high and low IQ's. These persons soon became all too ready to accept the test results as a final criterion by which a child's potential abilities might be determined once and for all.

Not only in its use of the IQ but in a number of other features did the new scale greatly surpass its predecessors. For the first time in the history of testing, a series of clear and well-organized instructions for administering and scoring the tests was provided. Previous versions of the Binet scale had included general instructions only. Such matters as the amount of urging permitted, whether or not questions might be repeated or supplementary questions asked, the exact criteria to be followed in scoring unusual or marginal responses were left, for the most part, to the judgment of the examiner. Terman not only covered these points in detail but also made it clear, as no previous constructor of tests had done, that deviations from the standard procedure which might

seem quite unimportant to the novice were likely to cause serious errors in the test results.

Attention was also given to the interest value of the separate test items and to matters of convenience in handling the results. Fractional scores were avoided by providing six items at each yearly age level. Each item thus counted for an even two months of mental age.⁴

The work of standardizing the scale occupied Terman and his associates for some five or six years. None of the items used were included merely because they had formed a part of the original scale as it was left by Binet. As a matter of fact, a number of these were discarded because, when tried out and carefully tested for their diagnostic value, they were found to be of little worth. Others, while retained in the scale, were considerably modified. Many completely new items were added. A valuable feature of the manual, especially for inexperienced workers, is a brief discussion, appended to the instructions for giving and scoring each of the test items, in which the underlying psychological factors which justify that item's inclusion in a scale designed to measure "general intelligence" are discussed in simple, nontechnical language.

Terman was also the first of the test constructors to realize clearly the importance of securing a representative sample of subjects for use in standardization. That he did not at that time fully appreciate the need for securing a sample of average variability as well as average developmental level is understandable enough, since few had stressed the matter at that comparatively early date in the history of testing. But he was keenly alive to the errors likely to result from irregularities in the definition of age⁵ or from a biased selection which might include too large a proportion of very bright or very dull children. He therefore used as subjects in his standardization group only those children who, at the time of testing, were within two months of their birthdays and who were attending schools located in what were judged to be "average" sections of American cities and towns.

The publication of the Stanford Revision marked the end of the

⁴ Except at the older ages, where tests were arranged for alternate years only.⁷ No tests were placed at Year XI, but eight were provided for Year XII, each of which counted for three months of mental age. Those at the later ages were similarly weighted according to the number of test items at each age and the length of the interval between the successive series.

⁵ When age is expressed only in units of an entire year, discrepancies large enough to cause appreciable errors in the averages of the successive yearly groups may appear if no attention is given to the matter. For example, the group classed as six-year-olds may have an average age of 6.8 years while those classed as seven-year-olds may average only 7.4 years. The age difference between the two is thus only 0.6 year instead of the full year presumed to exist. When such groups are used to determine the position of items in a year scale or for the establishment of normative standards by which other children are to be judged, irregularities are bound to occur.

initial period of experimentation and uncertainty. Once and for all, intelligence testing had been put on a firm basis. Terman's spectacular findings were duplicated by dozens of other experimenters who tried out the new instrument. Moreover, while Goddard had advocated the tests chiefly from the standpoint of their use in the diagnosis of mental deficiency, Terman laid equal stress upon the advantages to be obtained by their application to the understanding and guidance of normal and superior children.⁶

Evidence as to the general accuracy of the new scale mounted rapidly. It was found that in most cases the IQ's obtained by its use were reasonably consistent even when the tests were administered by different examiners and a considerable period of time had elapsed between the testings. Correlations with school success, with extent of general knowledge and information, and with other evidences of ability were appreciably higher than those obtained by the methods of appraisal previously used. Even those psychologists who, in the good old Wundtian tradition, had tended to look upon tests and testing with an air of lofty scorn were in many instances beginning to ask themselves whether this new device might not, after all, become a useful tool for fundamental research as well as in clinical practice. In his presidential address before the American Psychological Association, delivered in December, 1923, and published the following year, Terman gave a stoutly affirmative answer to this question. He pointed out that the tests are as well suited to the study of differences between groups as to that of differences between individuals, and that through their application to persons selected on the basis of some known but uncontrollable factor, many problems that had been thought not to be amenable to psychological investigation might be approached. He suggested that some of the discrepancies in the findings of different persons working on the same problem might be the unrecognized result of differences in the ability of the subjects used for experimentation and noted that such differences might be obviated if tests were used. He pointed out a wide variety of problems in the solution of which tests would aid.

THE STANFORD 1937 REVISION

For twenty-one years the Stanford 1916 Revision maintained the leading position among the intelligence tests used both in the United

⁶ Terman's lifelong interest in the child of superior intellectual gifts dates from his graduate years at Clark University. His Ph.D. thesis, published in 1906, was entitled, *Genius and stupidity; a study of some of the intellectual processes of seven "bright" and seven "stupid" boys*. It dealt with a comparison of the ways in which the subjects in the two contrasted groups responded to each of a series of specially devised intellectual tasks.

States and abroad. So great was the confidence engendered in this scale that many people came to identify a Stanford-Binet IQ with "intelligence" in much the same way as they would identify the reading on a balance with weight.

Nevertheless the 1916 scale was by no means perfect, and of this fact no one was more keenly aware than its author. In the first place, although tests were provided for the ages from three years to the level of "superior adult," the extreme ranges had been less thoroughly standardized and the tasks provided not always as well chosen as those for the ages from six to twelve. Particularly at the younger ages the calibration in whole-year intervals was much too coarse. The fact that only a single form of the test was available rendered it difficult to take care of the effects of practice when a retest was called for or to provide needed checks when the accuracy of a single test was questioned.

It was specifically to remedy these and other defects that a revision and extension of the 1916 scale was undertaken. Work on the new scales occupied a period of approximately ten years. The plan was undoubtedly the most ambitious ever undertaken for a similar purpose. The standardization group of about 3000 cases included subjects from all major areas of the United States. Both urban and rural children were included. In order to ensure that the sample would be as nearly as possible representative of the entire population in the United States, paternal occupations were ascertained in all cases, and the proportions of each occupational level as specified in the Minnesota Occupational Classification (Goodenough and Anderson, 1931) were matched with those reported in the United States Census of 1930. Two comparable forms of the scale were devised, each covering the age range from two years through adolescence and four levels of adult intelligence (average, somewhat superior, decidedly superior, and extremely superior).

The new scales, which appeared in 1937, have largely, though not completely, taken the place of the 1916 Revision. Some people still cling to the old form because of habit and because of the fact that it requires less time to administer. Its correlation with a single form of the new scales is about as high as that of the two new forms with each other for children of elementary school age, but for those under the age of six as well as for high school children and adults, the 1937 scales greatly surpass the earlier one.

A brief overview of the statistical findings for the new scales is given in the first four chapters of the Manual (Terman and Merrill, 1937); a more detailed presentation of the figures, with an account of the methods by which they were derived and a discussion of their meaning and significance, has been published by McNemar (1942). Many of the

points here brought out have far-reaching importance not only for the scales in question but for the construction and interpretation of other tests as well. These will be considered further in later sections of this book.

WORLD WAR I AND THE ADVENT OF GROUP TESTING

Immediately after the declaration of war with Germany in 1917, a special committee of the American Psychological Association was organized in order to answer the question: What can psychologists do to help win the war? It was decided that some device for aiding in the classification of men as to general ability and as to specific talents was urgently needed. Accordingly the committee, with R. M. Yerkes as chairman, set to work to build a scale of mental tests which could be used with the draft army both as a means of screening out those men whose intellectual limitations were so great that they would in all likelihood constitute more of a handicap than an asset to the fighting forces, and also, perhaps, as an aid in locating men who might be suitable candidates for officer training camps or for other positions of responsibility.

Obviously, the methods in use for testing individual children were too time consuming to be feasible. But the idea of group testing had already been born, though for the most part the method had been used only for the appraisal of educational skills of a rather mechanical nature. As early as 1908, Courtis had published his first series of objective tests of arithmetical processes, and others had followed with tests of other educational subjects. But the chief aid to the army psychologists was a group intelligence test not yet published upon which Arthur S. Otis was working. Otis immediately turned over all his data to the committee, and it was upon his model that the famous Army Alpha Test was constructed. The Beta Test, which was constructed for use with illiterates, proved to be rather less dependable than the Alpha, perhaps because the psychologists had less previous experimentation on which to build. However, it is noteworthy that up to the present time no one has succeeded in constructing a nonlanguage test which seems to be as useful as the verbal tests in appraising what is commonly thought of as "intelligence."

During the war the nature of the army tests was, of course, a military secret, but immediately after the close of hostilities many of the psychologists who had been working on these tests, as well as others who were impressed with their apparent simplicity and the reports of the

effectiveness with which they had served their purpose in the army, set to work to construct other group tests after the same general model. The list of these tests is far too long to reproduce here, nor would any useful purpose be served by doing so. The point that is important to note is the blind faith which many people displayed in the results obtained from them. The use of statistical methods as a means of demonstrating the worth of such devices was by this time getting into full swing, but few of the test makers had more than a rudimentary understanding of the procedures about which they talked so glibly. Discussions of the "reliability" and "validity" of the various new tests dominated the programs at psychological meetings or conferences on educational research, but many of the participants failed to realize that if two measures are correlated at all, a coefficient of almost any desired magnitude can be obtained by making the range of talent over which it is computed sufficiently long. "Mental ages" and "IQ's" obtained from half a dozen different group tests were joyfully computed and entered on children's permanent record cards by teachers and school principals with as much assurance as their grandfathers had placed in the skull maps drawn up by their favorite phrenologist. The decade of the 1920's was the heyday of the testing movement, the age of innocence when an IQ was an IQ and few ventured to doubt its omnipotence.

THE DEVELOPMENT OF NONVERBAL TESTS

Until 1921 there was no restriction upon the volume of immigration to the United States from the overcrowded nations of Europe. Unskilled labor was needed in the factories of the East and on the ranches of the Western states. In many European countries food was scarce and jobs ill paid and hard to secure. Fabulous stories were told of conditions in America, where a man might earn as much in a day as he was paid for a week of hard labor in the "old country." The steerage of every ship leaving Europe for America was packed with immigrants of all types. Some of them had the makings of good sturdy citizens; others were physically or mentally incapable of self-support even under the most favorable conditions. Upon arrival at Ellis Island all were presumably examined by physicians whose task it was to weed out those who were physically or mentally defective. But the physicians were often not highly competent, and the amount of time available for the examinations was in any case far too short for an adequate appraisal; hence the number of mentally defective cases admitted was disturbingly large. With growing concern about school retardation came the recognition that a disproportionately large number of the retarded children were

immigrants or the children of immigrants. While a part of the educational difficulties of these children could be traced to imperfect knowledge of English, it soon became apparent that this was not the whole story. Moreover, as social work became better organized and social workers better trained, the feeble-minded immigrant became an ever-recurrent social problem as well. Evidently the screening methods in use at Ellis Island were ineffective. Additional safeguards were called for.

It was specifically to meet this need that Knox devised a series of tests which required no use of spoken language. The tests were not well standardized, but they proved to be more effective in selecting at least the grosser cases of mental defect than had the hasty medical inspection previously depended upon. A number of the tests used by Knox were subsequently incorporated into other scales which were more adequately standardized. Knox published a report of his work at Ellis Island in 1914. Three years later, in 1917, Pintner and Paterson brought out the first well-standardized scale of nonverbal or, as they now came to be called, "performance" tests. Other scales of the same general sort soon followed. Of those in use today, the Arthur Point Scale (1930) and the Cornell-Coxe series (1934) are among the best known.

It is unfortunate that up to the present time no really conclusive analysis has been made of the psychological differences between tests of the Binet type and those commonly used in the so-called "performance" scales which are typically made up of form boards, picture puzzles, mazes, and so on. Superficial examination of the nature of the tests and of the results which have been reported from their use shows certain facts to be reasonably well established. The performance tests thus far developed are of relatively little use in determining mental differences among normal adolescents or adults. For representative groups of English-speaking children below the age of eleven or twelve years whose chronological ages do not differ by more than a year, correlations with the Stanford-Binet test have as a rule not exceeded .75. In other words, the error of estimating a child's most probable score on one of these scales from a knowledge of his standing on the other would be reduced about a third below that to be expected on the basis of sheer guess. (See Chapter 11.) Although such an improvement is distinctly worth while, it leaves room for many points of difference between the two measurements. It certainly does not warrant the assumption that one may be taken as the equivalent of the other.

The common practice of scoring performance tests on the basis of time and number of errors obviously places a premium upon speed. The extent to which speed should be regarded as an attribute of intelligence is a moot question about which there have been many opinions and a

fair amount of experimentation. That the experimental results have not led to uniform conclusions may perhaps be traced, at least in part, to the fact that speed is not an abstract characteristic, independent of the acts in which it is manifested. Its significance probably varies with the type of performance and possibly with level of performance as well. Speed, moreover, appears to be more easily affected by differences in interest and effort than is the quality of performance considered without reference to speed.

EDUCATIONAL TESTS

The use of standard tests to measure proficiency in the various subjects of the school curriculum dates back to the very early days of the testing movement. School examinations devised by teachers or other school authorities have, of course, a still longer history. The advantages of the standardized test over lists of questions made out by teachers are numerous. Among the most important may be noted those resulting from the more careful selection and formulation of the items comprising the tests. As a result of the establishment of comprehensive standards of reference, certain general comparisons are made possible. The performance of individuals or groups may be compared with that of larger and presumably more representative populations than are immediately available to any one teacher. Provided the nature and limitations of such comparisons are clearly understood, much valuable information may be had in this way. But educational tests, like intelligence tests, have often been misused by the uninformed. Of this, more later.

The earliest of the educational tests to be standardized had to do with the more mechanical features of the curriculum, such as arithmetic computation, spelling, speed of reading. These were followed by a series of product scales, devised by Thorndike and his students and scaled on the Fullerton and Cattell principle that differences equally often noticed are equal.⁷ By the use of this method, scale values are derived for each of a series of sample products, such as specimens of handwriting or drawing, which are then arranged in order of merit. In using the scale, the sample to be scored is compared with each of the standardized samples in turn until one of roughly corresponding merit is found. The scale value of this standard sample is then assigned as the child's score. Intermediate scores may also be given in cases where the child's product is thought to be better than one of the standards but not quite so good as the one following.

Tests of knowledge or information quickly followed. These were

⁷ See Chapter 22.

both specific, dealing with such topics as history, geography, literature, general science, and so on, and more general, covering a wide range of topics and as a rule including not only items commonly learned at school but also such extracurricular activities as games and sports and practical affairs of everyday life. Tests of the latter kind are usually called tests of "general" information. In most cases, both these and the more specific tests dealing with individual subjects of the curriculum are arranged in the multiple-choice form and scored on a point system. By reference to appropriate tables of standards, the point scores may then be converted into various interpretative units, such as educational ages—corresponding to the now familiar mental ages—educational quotients, percentile ranks, standard scores,⁸ and so on.

As with intelligence tests, the early educational tests commonly dealt with single topics only. It was not long, however, before the idea of developing complete batteries of tests covering all the major subjects usually taught in a given grade occurred to the test makers, and by the time the use of group intelligence tests had become common a number of such batteries had appeared. By the use of these batteries it became possible to note where a given child's chief educational weaknesses lay and thus direct his school training more definitely in accordance with his individual needs. It also became possible to compare his general educational standing with his performance on intelligence tests and thus to judge whether or not his schoolwork was of a quality comparable to his ability. But such a comparison, to be valid, is not a simple matter of comparing educational ages or educational quotients with mental ages or IQ's as Franzen (1920) originally recommended and as many poorly informed educators are still doing in spite of repeated demonstrations of the statistical fallacies involved.⁹

The next major step in the development of educational tests consisted in the finer classification of educational measures according to the type of psychological function involved. Reading tests were differentiated into tests of speed and tests of comprehension, and the latter were further classified according to whether the comprehension involved small units such as single words or short sentences, or longer units such as paragraphs or short essays. Tests of arithmetic reasoning were distinguished from tests of computational processes. Tests of speed were distinguished from tests of accuracy; tests of factual knowledge were differentiated from tests of the ability to apply these facts in the explanation of scientific and social phenomena. All this led directly to the development of diagnostic tests which had as their aim not merely the classification of

⁸ See Chapter 13.

⁹ See Chapter 22.

children according to level of achievement but, what is vastly more important, the tracing of a particular deficiency shown by a child back to a more fundamental level and thus indicating the kind of corrective measures most likely to be helpful. A large number of such devices for the diagnosis of reading difficulties are now on the market. Many of them have been developed on the basis of some special theory of a common causative factor underlying most cases of poor reading. Their usefulness is therefore dependent, at least in part, on the soundness of the theory in question. Others are more eclectic in nature, built chiefly upon the idea that poor reading necessarily involves poor reading habits and seeking primarily to ascertain more exactly the nature of these habits. Although reading continues to be the area toward which the makers of diagnostic tests have directed most of their efforts, some attempts in this direction have been made in most of the other subject fields, especially arithmetic and handwriting.

Measures of educational achievement are likely to have some predictive value as well, inasmuch as the person who has already demonstrated superior ability up to a certain level is, on the average, more likely to be able to progress beyond that level than is another whose achievement has hitherto been poor. Obviously, however, a good deal of wasted time and effort might have been saved if some method had been available for predicting success or failure in advance of actual trial. Particularly is this the case at the high school and college levels where a choice is to be made in respect to the subject matter covered. "Aptitude tests" of various kinds have accordingly been devised. Some of these are designed as tests of aptitude for certain vocations; others, more specific in nature, are tests of aptitude for certain specified educational subjects which may or may not be directed toward specific occupational fields. There are, for example, a number of more or less well-standardized tests of aptitude for learning foreign languages, mechanical drawing, advanced mathematics, or science, as well as the more direct vocational tests of clerical aptitude, mechanical aptitude, and the like.

For use at the younger ages where education is still general rather than specialized and all children are expected to learn at least the rudiments of each of the various subjects included in the curriculum, a somewhat different kind of aptitude test has become increasingly popular during recent years. This is the so-called "readiness" test, constructed on the premise that neither the chronological nor even the mental age of a child furnishes the best possible evidence as to whether he has reached a stage of development at which he is "ready" to learn a particular skill, such as beginning reading. "Reading readiness" tests have been found decidedly useful in helping to decide whether or not

it is wise to promote children from kindergarten into first grade, particularly when their mental or chronological ages are near the boundary line at which such promotion is customarily given or withheld. Tests of "arithmetic readiness" and the like have also been devised, but since the subject of paramount importance in the first grade is unquestionably reading, and since, once the elementary school has been entered, the time of taking up the various subjects is commonly determined by school practice rather than by the readiness or unreadiness of the individual child, the latter tests have been of less practical usefulness.

TESTS OF MOTOR AND PERCEPTUAL ABILITIES

The success of the tests of intelligence and of educational achievement led a number of people, particularly those interested in physical education, to wonder whether or not it might be possible to measure general motor ability in a similar manner. It was noted that some children seem to be much more adept in handling their bodies than others; that they learn to dance, swim, and skate with relative ease while others require more time and never attain the same degree of proficiency; that some are deft in their handling of tools and other objects while others seem all thumbs. That such a general characteristic as "motor ability" might underlie all these differences was a natural hypothesis, and attempts to devise a scale for its measurement were soon forthcoming.

It was found, however, that unlike the tests of "general intelligence," which commonly tend to show a good deal of correspondence in results when applied to the same subjects, even though they bear but slight superficial resemblance to each other, the various tests designed to measure motor ability are likely to yield divergent results unless they are quite similar in nature. Particularly it has been found that dexterity of hand bears little or no relation to motor skills involving the larger muscles of legs and trunk. Moreover, even when the same general muscle groups are involved, the use of a different criterion for judging the quality of performance often results in a very different classification of the same performer, or even of the same performance. The boy who wins the fifty-yard dash is not necessarily the one who comes out best in a two-mile race; the fastest swimmer is not always the one who shows the best form; the football champion may be a poor tennis player. Either because there is no agreement as to the fundamental nature of "motor ability" or because motor abilities are actually so distinct from each other that there is no warrant for attempting to throw them into a single category, no really successful scale for measuring general motor ability, comparable to tests of general intelligence, has as yet been derived.

There are, however, a very large number of highly reliable tests for measuring specific motor skills, both manual skills of various kinds and large-muscle skills. While it is true that there is a general improvement in most types of motor skill with advancing age, this does not necessarily mean that at any given age a child who is advanced in one skill will also be above the average of his age group in other skills.¹⁰

Perhaps the most notable of the attempts to develop a general scale of motor abilities is that by Oseretzky (1931). The unique feature of Oseretzky's approach lies in the fact that he arranged his test items in the form of a year-scale after the manner of a Binet test. The method of administering and scoring is also similar to that of the Binet. Because of the fact that motor abilities do show advancement with age, the procedure has a certain plausibility, but growth with age is not the only criterion that must be met if the year-scale method is to be valid. There must also be internal consistency of the items, which may be shown either by correlation of each one with a common criterion or by the correlation of each separately with the sum of all the others when the factor of age is held constant. Oseretzky's tests have never been subjected to really critical analysis by a competent statistician. Whether or not they would stand up under such treatment is uncertain.

Many batteries of motor tests, with the results of each reduced to similar units but without attempt to combine them into a single score, have been made up, and there have also been several fairly successful tests of gross motor ability embodying a number of somewhat similar motor tasks intended for use only over a limited age range, such as high school or college.

An aspect of motor ability that has been of special psychological interest is lateral dominance. That nearly all people use one hand in preference to the other for most skilled acts is well known. That the extent of this tendency varies all the way from near or complete ambidexterity to a very marked hand preference is likewise a matter of common observation. A number of reasonably dependable scales for measuring the extent of hand preference have been devised, of which one of the most recent and dependable is that by Johnson (1936, 1937). Various methods, most of them dependent upon measurement of the action current in the nerves of the two arms, have also been proposed for determining the "native" laterality in cases where laterality prefer-

¹⁰ Mentally defective children, especially those of the lower grades such as idiots and imbeciles, are also, as a rule, retarded in practically all lines of motor development. But if the grossly defective cases are excluded, neither the correlation between different motor skills nor that between motor skills and intelligence is sufficiently high to be of much practical use for predicting one from the other.

ences are believed to have been altered in early childhood. The significance of these measures is somewhat questionable since their validity as indicators of original laterality preference rests upon assumptions that have never been satisfactorily established.¹¹ Until such evidence is forthcoming, it appears safer to stick to the overt indications of present laterality tendencies which involve no assumptions as to how the preferences originated.

Eye dominance can be satisfactorily measured by means of a manoptoscope,* a device originally developed by Parson (1924) and subsequently simplified and improved by Miles (1930) and others. No very satisfactory way of measuring foot dominance has been devised, though it is probable that some degree of laterality tendency is also present in the use of the feet and legs.

TESTS FOR INDUSTRIAL SELECTION AND VOCATIONAL GUIDANCE

Although the idea of devising tests that may aid in selecting workers for particular industries or in advising young people regarding the type of work for which they are best fitted arose early, the great impetus for the development of such tests came, like the development of group intelligence tests, as a result of the psychological studies in World War I. That the increased interest in the two came at about the same time is not a matter of accident, for as long as the trades tests had to be administered individually, not many firms were willing to undergo the necessary expense, nor were many school systems able to provide such vocational service for more than a small proportion of their students.

It would not be profitable to attempt to give a description of the great variety of trades tests and tests that have been developed in the

¹¹ Whether or not such a thing as "native" laterality preference, independent of training and habit, actually exists is not certainly known, though the fact that the great majority of the races of mankind appear to be more or less right-handed lends plausibility to the hypothesis that such is the case. The right eye is also dominant over the left in about two thirds of all cases. Metfessel and Warren (1934) have shown that when a subject with well-established lateral dominance, in whom there is no reason to believe that any attempt to alter the tendency as overtly displayed was ever made, is asked to reach with both hands simultaneously toward an object placed in the median line from the body, the action currents usually start first in the arm on the nonpreferred rather than the preferred side as had been assumed by those using this method for the clinical detection of "native" dominance. The matter is of considerable practical as well as scientific importance because of its relation to theories of the causes of stuttering and the attempts of many clinicians to correct stuttering by training the stutterer to use the nonpreferred hand for all major activities upon the assumption that hand dominance had been changed by parental or other authority during early childhood.

* See Figs. 30, 31.

course of the past thirty years for the vocational guidance of young persons who have not yet decided upon their lifework. We may note, however, certain main distinctions important to keep in mind when dealing with tests of this kind. For the most part, these distinctions center around two questions: (1) For what purpose is the test to be used? (2) Is technical ability or the personality characteristics and type of interests of the subject held to be the more important factor in success, granting that both play some part in the matter?

In respect to the first question, one must draw a sharp distinction between tests intended merely for selecting individuals to fill a particular job, as the employment manager in an industrial concern has to do, and the vocational guidance of individuals who have not yet decided upon the kind of job they should seek or the field for which they should prepare themselves. It is obvious that of the two problems the former is much the simpler.¹² That personality characteristics as well as technical skill should be taken into account in deciding upon the kind of work in which an individual is most likely to win happiness and success is a concept that was recognized in a general way at a fairly early date but did not pass beyond the stage of poorly standardized rating scales, personal judgments, and occasional questionnaires until 1922, when Max Freyd published his Occupational Interest Blank. Although prepared in the form of a questionnaire dealing with interests of many kinds, the questions were actually selected in such a way that differences in many aspects of personality were revealed by them. Freyd had his questionnaire filled out by persons actually engaged in several different professional fields and was able to show that members of these professions were characterized by rather distinctive patterns of interest which extended far beyond those matters directly concerned with their chosen lines of work. Freyd's blank was revised by Cowdery and later by Strong, who developed a multiple scoring key by means of which a student's "interest patterns" could be classified with respect to those found characteristic of each of a large number of occupations. Since his original publication in 1926, Strong has revised his questionnaire several times and added scoring keys for many occupations not originally included. A second blank for use with women (the earlier forms were intended only for use with men), with appropriate scoring keys, has also been devised. The two questionnaires are similar but not identical, and the scoring of the various items is not always the same for both sexes. In his last major publication (1943) Strong makes a careful analysis of the nature and significance of sex differences in expressed interests, and presents a special key for

¹² These questions are taken up in more detail in Chapter 28.

scoring the interest questionnaire on the basis of the masculinity or femininity¹³ of the interests claimed.

THE APPRAISAL OF "PERSONALITY"

That ability is not an infallible indicator of performance and that conduct depends upon many factors with which reason and judgment have little to do has been recognized for centuries. As a matter of fact, the concept of the "personality trait" as a relatively fixed and invariant attribute of the individual is even today accepted by many psychologists without serious question. Perhaps there is no other aspect of mental measurement about which such wide discrepancies between theory and practice are found, or it may be that there is no other area in which so much overt activity is accompanied by so little thought!

Before the end of the last century a number of persons, including Binet, had experimented with methods which, it was hoped, would throw some light upon the factors underlying the differences in behavior of persons who, as far as intellectual ability of the more abstract type was concerned, seemed equally well endowed. Among other devices, Binet (1898) designed a test of "morality" in which children were asked what they thought should be done if a child accidentally broke a violin belonging to a friend. Responses were classified according to a prepared list of alternatives.¹⁴ Tests similar to this were included as parts of a number of group tests of "personality" that appeared during the decade of the 1920's, but so little relationship was found between the responses given by children and their actual behavior that the popularity of these "ethical judgment tests," as they were called, soon declined.

During World War I, R. S. Woodworth devised a questionnaire which he called a "Personal Data Sheet," the object of which was to screen out men having such marked neurotic tendencies as to unfit them for military service. The device proved so useful¹⁵ that when it became available at the end of the war, its possibilities for civilian use, particularly with school children, were immediately noted. In its original form the ques-

¹³ See Chapter 35.

¹⁴ It may be noted that for young children, questions of this kind have been found useful indicators of intellectual development, inasmuch as the conventional answer demands that the child has reached a level of understanding at which the abstract principle of *intent* is regarded as of greater importance than the concrete act or event. Both the 1916 and the 1937 Stanford Revisions of the Binet scale include one or more items of this kind.

¹⁵ It should be unnecessary to point out that the questionnaire was used only for preliminary sifting. The final decision as to whether or not a man was suitable for army service was based upon an individual examination by a psychiatrist whenever there was doubt.

tionnaire was unsuited for use with children because many of the questions asked would not apply to them. With some modifications, however, it was possible to adapt it to the child level. The general principles of this questionnaire, as well as a considerable number of the original questions, have been embodied in a large number of modern scales. Originally described as questionnaires for measuring "emotional stability," they are now more often referred to as "adjustment questionnaires" or "personality inventories." By the use of multiple scoring keys, various areas of adjustment are delimited.¹⁶ In this way a more precise analysis of the particular kinds of difficulty which an individual is experiencing can presumably be made.

The great advantage as well as the greatest hazard in the use of the adjustment questionnaire for appraising personality tendencies lies in its directness. It asks for personal experiences and attitudes, opinions, beliefs, interests. Since matters of this kind are not, as a rule, open to general observation, the only person who can give information about them is the subject himself. But since a large proportion of the items usually included in these questionnaires have to do with matters in which one type of response is generally recognized as the socially desirable one,¹⁷ the question of frankness in response inevitably arises. The proponents of this method have attempted to solve this difficulty in a number of ways. In the first place they have pointed out that the responses to these questions are not looked upon as objective facts but as claims. If regarded as claims, their significance as indicators of personal adjustment can be determined in the same way as other responses by a combination of the methods of internal consistency and comparison with an outside criterion. If all persons responded with the same degree of honesty, this argument would unquestionably hold good and a scale of which the items were selected on this basis would inevitably "work" with the majority of persons not too unlike those used for standardizing.

However, in spite of its continued popularity, the personality inventory has been the subject of severe criticism from a number of quarters. To some extent, at least, this criticism is justified. Because of the difficulty of finding a dependable outside criterion with which the results can be compared, the constructors of these questionnaires have undoubtedly leaned too heavily upon armchair judgment in deciding whether a nega-

¹⁶ For example, adjustment to home and family, adjustment to school, hypochondriacal tendencies, feelings of inferiority, social adjustment, and so on.

¹⁷ Examples:

Are you happy most of the time?	Yes	No
Do your teachers usually treat you fairly?	Yes	No
Do you think you have as many friends as other people have?	Yes	No

tive or a positive response to a given question is to be preferred, as well as in deciding which of the questions should be grouped together to form subscales. The lack of an outside criterion has also led to too great dependence upon the factors of internal consistency of the items and of similarity of total scores attained on successive repetitions of the same scale or on different forms of a scale made up of closely similar items. But neither of these factors is a sufficient guide to interpreting results. It is a statistical truism that a test may have high "reliability" and low "validity" with respect to the purpose for which it was designed. Although a number of the inventories designed for children meet the test of consistency reasonably well, the evidence for the validity of the interpretations specified by their authors is less convincing.¹⁸

It is enlightening to encourage a child who is sufficiently at his ease to express himself freely to "think aloud" while he fills out a questionnaire of this kind. Here are a few examples from my own experience.

Subject: An eight-year-old boy of superior intelligence. Apparently happy and well adjusted. His family has always placed a good deal of stress upon truthfulness and honesty.

Q. Are you afraid of the dark?

Child. Well now, I don't know what to say to that one. Mostly I'm not afraid at all. But the other night, well you see I was coming home from John's house—I'd been to supper there and it was dark and you know that dark alley I have to go by—well you see I thought I saw somebody there and I thought about that boy in the paper [a recent kidnapping case] and did I run! Well—[returning to paper] I s'pose that means if you're *ever* afraid. I guess I better mark it "yes."

With a vigor born of the difficulty of his decision he drew two heavy lines under the word "yes" and turned to the next question. This boy's total score fell in the group classed as *very unsatisfactory*, but no one who had listened to his comments while filling out the blank could doubt that most of his "undesirable" responses arose from a high regard for the truth coupled with a straightforward and comparatively unemotional acceptance of his own shortcomings as facts to be dealt with but not worried about.

Subject: A girl of ten, coming from a home where there was much friction between the parents and lack of agreement as to management of the children.

¹⁸ In a recent survey of the literature, Ellis (1946) presents a rather scathing criticism of a number of the most widely used tests of this kind. His conclusion that few if any of these inventories have sufficient value with respect to their designated purposes to warrant their use may, however, be questioned on the grounds that his basis of comparison is not always completely fair (as when he condemns a scale designed to measure introversion on the grounds that it does not distinguish between delinquents and nondelinquents), nor is his practice of classifying complex scales into single broad categories of satisfactory or unsatisfactory one that is usually to be recommended.

Q. Do your parents usually treat you right?

The child read the question, hesitated, and then drew a rather faint line under the word "yes." She made no comment at the time but a few minutes later she remarked, "I don't think it's nice for children to tell things about their homes, do you? You know Vera Andrews? Well, she told me her mother always yells at her when—you know, if she don't pick up her things or if she breaks a dish or something. I told her she shouldn't say things like that about her family."

Subject: A boy of seventeen, failing in practically all his high school courses, low average mentality. His parents were much disturbed over the likelihood that he would not be recommended for admission to college. A good deal of pressure was being used to try to get the boy to do better work.

Q. Do you usually get good marks in school?

Answer: Yes.

Q. Do you often find your lessons hard to do?

Answer: No.

Q. Would you like to leave school and go to work?

Answer: Yes.

With the possible exception of the last item, there was nothing in the boy's responses to any of the questions regarding school that would suggest difficulty in that area. But practically all the questions having to do with his state of health were answered in the "undesirable" manner. He claimed to have poor vision, frequent headaches, to be easily fatigued, to sleep poorly, and to have many bad dreams. A casual question brought a flood of elaborations. He was, he said, always getting something the matter with him. Always catching cold. "Of course," he added, "I've had to miss a lot of school. That's why I'm behind." He was reminded of his statement that he usually got good marks. "Well—" he stammered, "I—that is, when I'm feeling good, I do. But when I have such an awful cold or a headache or something I just can't think. Last year I was awful sick and I had to stay home so long I didn't pass. But I couldn't help that could I?"

Actually, however, a recent physical examination showed the lad to be in excellent health and his attendance record revealed only an average number of absences. The "long sickness" of the previous year had been a mild case of influenza which had kept him at home for exactly a week. The claim of poor health in his case seemed to be not so much a hypochondriacal tendency as a rather transparent defense mechanism by means of which the boy was protecting his pride. It enabled him to claim good marks "if he was feeling well" since by a simple reversal of reasoning he could persuade himself that if his marks were not good it must be that he was *not* feeling well.

While cases such as these are not very usual, they still occur often enough to point to the need for caution in interpreting material of this kind. There is some reason for thinking that such inventories have greater

diagnostic value for adults than they have for children. If this is the case, it may indicate that adults interpret such questions in a more uniform and direct manner than children do. It is also possible that the questionnaires intended for adults have been more carefully constructed and standardized than those used with children.

Very different from the personality questionnaire but based upon much the same general principle are the various "performance tests" designed to measure specific aspects of personality and conduct by means of sampling. Voelker (1921) was one of the first to prepare a fairly comprehensive battery of such tests. The method consists, essentially, of setting up a series of concrete situations in which opportunity is provided for the occurrence of the kind of behavior in which the experimenter is interested, together with concealed methods for detecting it. Cheating, stealing, lying, and a number of other forms of misconduct have been studied in this way. The most extensive use of tests of the performance type was made by Hartshorne and May and their associates (1928, 1929, 1930). Their results ran so strongly counter to the generally accepted idea that conduct is merely the outward expression of generalized personality traits such as honesty or dishonesty, generosity or selfishness, and the like, that for some years but little further attempt was made to develop tests by which such behavior might be predicted. Even up to the present day, most people have preferred to steer clear of these topics or have employed other methods for studying them.

The study of attitudes toward various social institutions, ethnic, religious, and political groups, and similar matters has been most completely developed by Thurstone and his associates at the University of Chicago. Most of the statistical methods commonly used in the construction of these scales were worked out by Thurstone, but a number of other people have developed various technical modifications and supplementary devices for use in special cases. Although by far the greater number of studies in attitude measurement have used questionnaires especially constructed for the investigation of some particular question, a large number of formally constructed scales of this kind are available commercially. The same may be said with regard to the measurement of interests (other than vocational interests), beliefs, and opinions. The public opinion poll, of which more will be said in a later chapter, is a more recently developed variant in this line.

"PROJECTIVE" METHODS FOR ASSAYING THE PERSONALITY

The fact that each individual lives in a "private world" of his own making (Frank, 1939), that he has aspirations, attitudes, emotional

drives, fears, and beliefs which are not apparent to an outsider and as a rule not clearly sensed by the individual himself is a fact that has become increasingly apparent to those who have been attempting to probe beneath the surface in their studies of the mechanisms of human behavior. That such methods as have been described in the preceding sections are incapable of uncovering these hidden depths is generally admitted. Neither direct sampling of abilities and behaviors nor the substitution therefor of verbal symbols in the form of questionnaires, rating scales, attitude scales, and so on, can reveal more than fleeting and often distorted glimpses of the springs of human action. Two persons of similar ability and equal opportunities may nevertheless behave very differently. This is a truism, but why they do so is a question that up to now has not been satisfactorily answered.

The series of devices which have been developed in the past few years, largely for use with children, and which are generally grouped under the heading of "projective methods," represent an attempt to assay the hidden emotional urges which are the starting points of overt behavior. The underlying premise of the devices used is that the child's inner urges must find an outlet, and when the normal or usual outlet is blocked off through social pressures, through fear of punishment, or for other causes, the feeling will be "projected" onto some other object or situation. Thus the child who feels resentment toward his parents or jealousy of a brother or sister but who, either through fear of punishment or for other reasons is unable or unwilling to vent his feelings directly upon the individuals concerned, will "project" them upon some other object which is taken as a symbol.

The number and variety of devices which have been used in "projective" experiments are far too great to be enumerated here. Dolls representing the various members of the child's family, fragile objects such as toy balloons which the child is permitted or even encouraged to break, smeary substances such as dough or cold cream, materials for artistic or dramatic expression such as paints or toy theaters are examples. The idea underlying these methods is unquestionably a promising one, but it is unfortunate that in too many of the studies thus far carried out, interpretations have been made on so shallow and superficial a level that most, if not all, of the potential value of the concept has been lost. The difficulty seems to lie in the failure of the persons concerned to realize the fundamental difference between a sign and a sample. The methods which we have discussed up to this point are for the most part sampling devices. Samples are taken of the child's ability to solve difficult problems of various kinds, to remember different kinds of material, to react quickly

in response to a signal, to run, jump, lift weights. Samples of the opinions of other persons regarding his behavior, or his own expressed claims about his interests, feelings, or beliefs are secured, and within the limits of sampling error, these are assumed to be representative of the larger area of similar abilities, behaviors, and claims from which the samples have presumably been drawn. But the essential feature of the "projective" techniques lies in the fact that the behavior shown is not to be looked upon as a sample but as a sign. And it is of the very nature of a sign that its overt character does not necessarily resemble the thing signified.

An outstanding exception to the types of projective method just mentioned is the Rorschach test. In this test the subject is required to interpret a standard series of inkblots in terms of what the patterns of line and form suggest to him. In contrast to the procedure followed in most projective methods, his responses are not interpreted in terms of their surface characteristics. The test is scored on the basis of four general aspects of the responses: (1) their *location*, in which account is taken of his tendency to base his interpretations upon large or small areas as well as a number of other aspects having to do with the part of the blot to which response is made; (2) the *determinants* of the response, under which term are included those characteristics of the blot which seem to be mainly responsible in determining his perceptions, such as form, shading, color,¹⁹ and so on; (3) the *content* of the responses, under which are classified the general kind of objects named; and (4) the degree of *originality* of the responses, which is determined by a comparison of the subject's associations with those of persons in general.²⁰ The *time required for responding*, the *number of responses given*, and various measures of the *pattern of response* obtained by finding the ratio between the frequencies of certain specified types of response by means of standard formulas are also determined. From all this information various conclusions regarding the abilities, interests, and personality characteristics of the individual are reached. A practiced Rorschach examiner commonly presents his findings in the form of a personality sketch, phrased in descriptive rather than numerical terms. A number of such sketches may be drawn up on the basis of the Rorschach responses alone by a person

¹⁹ Some of the blots in the series are done in black ink only; others include both black and colors.

²⁰ Tables, showing the distribution of responses of various groups of persons who have been given the test, are available for comparison. However, these tables are less comprehensive than would be desirable, and the ages, educational level, and other characteristics of the subjects upon whom they are based have not always been clearly indicated.

skilled in the analysis of such data but who has never seen the subjects,²¹ and the sketches may then be presented to a third person familiar with the subjects but not necessarily with the Rorschach method as such. This person is given a list of the persons to whom the sketches apply and is asked to match names and descriptions. In most of the experiments of this kind that have been reported, success in matching is better than chance would lead one to expect, a fact which lends some support to the claims made for the method.

Further discussion of this method will be found in Chapter 27. For the present we may note only that in many respects it stands in a class by itself. More than any other testing procedure that has been developed up to the present time, with the possible exception of those used by Wolff (1943, 1946) in his so-called "Depth Psychology," this is a method which makes use of signs and not of samples.²² And because the sign is a molar and not a molecular concept, it has special appeal for the increasingly large number of psychologists who look upon personality as a molar affair and not simply a bundle of separate and unorganized traits or tendencies.

The major difficulty with the Rorschach method at present lies in the fact that "signs" have been identified and their significance proclaimed by so many enthusiastic dabblers in the method who are lacking in basic scientific training and who fail to distinguish between coincidence and correlation. In their zeal for discovery they fail to verify what they have found, and as a result the literature has become flooded with reports of such doubtful validity that many conservative psychologists are inclined to wonder whether the whole procedure might not better be classed as a cult rather than as a scientific method. Nevertheless there are a goodly number of well-trained persons who see in the method the possibilities of an approach to the study of the foundations of character that cannot be made through the simpler and more direct methods based upon sampling. From the investigations of these persons a framework of established fact is gradually being built which in time may develop into a completed structure.

²¹ This procedure is known as the method of *blind analysis*. The Rorschach tests are given by one person who records the responses only. The record blanks are then sent to a second Rorschach expert who prepares the character sketches with no information regarding the subjects beyond that included in the record blanks. The matching is done by a third person, not necessarily a Rorschach expert, but who is well acquainted with all the subjects.

²² A further exception should be made with respect to such procedures as the method of free association used by Wyman (1925) for the study of certain types of interests and later by Goodenough (1942, 1946) for the study of mental masculinity or femininity and other personality traits. But the Wyman method of scoring was never published, and the Goodenough scoring keys have not as yet been published, though it is expected that they will be in the near future.

ADVANCES IN STATISTICAL METHOD

At the time of the first translations and adaptations of the Binet tests for use in America, the application of statistical methods to problems of tests and testing was neither well understood nor thoroughly appreciated. Galton, to be sure, had from the beginning emphasized the need for mathematical treatment of psychological data. Later, under the influence of Pearson, Yule, Thorndike, Brown, Elderton, Spearman, Fisher, "Student," and others, the conviction became firmly rooted that statistical analysis is not merely a helpful device for determining the adequacy of tests but a prime necessity in their construction. Statisticians, both little and great, whose major interest centered around the field of mental and social measurements increased in number so rapidly that merely a list of those best known would make up a long paragraph. New techniques have been developed both for the solution of new problems and for the refinement of old methods. Much attention has been given to the statistics of probability. Determination of the likelihood that an obtained result will be duplicated, as far as its general direction or tenor is concerned, has become practically a *sine qua non* in all studies involving psychological tests or measurements. Attempts at statistical breakdowns of psychological measuring devices into more homogeneous elements or "factors" were begun by Spearman in connection with his "two-factor theory"²³ as early as 1904 and were later elaborated by Kelley, Holzinger, and others. The most elaborate and extensive work in this field has been done by Thurstone, who, in addition to many contributions to the technique of "factor analysis," as it is now called, has utilized the method in the development of a series of tests designed to be more nearly "pure" or homogeneous measures of such talents as verbal facility, perceptual abilities, memory, and inductive ability. To these Thurstone gives the name of "primary abilities" because he believes them to be relatively independent of each other and to constitute some of the basic elements of which abstract intelligence is made up.

Other important developments of the past few years include the derivation of special methods for use when working with small samples, the analysis of variance and covariance as developed by R. A. Fisher and his associates, together with emphasis upon the advantages of designing an experiment in advance rather than depending upon later analysis of data collected without formal plan. Greater attention has also been given to questions of sampling, and a number of special devices and procedures for testing the character of a sample have been developed. The importance of these methods can hardly be overestimated since, as

²³ See Chapter 19.

has previously been pointed out, most of our procedures for appraising the characteristics of individuals or of groups are sampling methods, and their usefulness is contingent, first, upon the representativeness of the sample of items with reference to the characteristic to be appraised and, second, upon the representativeness of the group of subjects whose performance on these items is to be taken as a standard with which that of other groups is to be compared.

Although the modern emphasis upon the use of statistics in the evaluation of tests and of the results obtained by testing has unquestionably led to a tremendous improvement in procedures and to a clarification of many issues that were formerly vague or unrecognized, it has nevertheless not always proved itself to be an unmixed blessing. The modern university student of psychology, education, or sociology takes one or more courses in statistical method as a matter of curricular requirement. He learns statistical formulas and perhaps their mathematical derivation as well. He learns by rote how to find means and standard deviations, compute correlations, make factor analyses. And in many instances he also learns by rote to "interpret" the results of these computations. He feels assured that those findings which reach certain arbitrarily set levels of "significance" are immutably established, that chance stops short at those points. But if his figures fail to reach this magic criterion, be it only by the third decimal place, he sweeps them aside as "not significant." He computes the standard error of a single score and painstakingly records it, but a "gain" or "loss" as determined by the difference between two such fallible scores is treated as if it were free from errors of measurement. If, as is unfortunately all too common even today, he has been taught that some definite IQ level, say 70, marks the border line between the "normal" and the "feeble-minded," he is ready to consign a child with an IQ of 69 to an institution without delay, whereas another whose test performance places him at 71 is expected to comport himself as a normal individual. A child of seven with a mental age of ten and an educational standing equal to that of the average child of nine is said to be "not working up to capacity," while another with a chronological age of twelve and a mental age of nine whose educational achievement is likewise equal to that of the average nine-year-old is said to be "working up to his capacity" in spite of the fact that if both entered school at the age of six, the first child has accomplished the work of three years in the course of a single year, whereas the second has taken six years to do the work of three. The fallacy, of course, lies in the different location of the zero point in the two measurements. These are but a few examples of the erroneous interpretations of

statistical results by those who have learned procedures but are ignorant of the basic assumptions upon which these procedures are based.

We have come a long way since 1905, when Binet first showed that mental tests can be useful. We have learned much, both on the theoretical and on the technical side. From a state of widespread skepticism, both scientists and laymen have for the most part grown to accept the idea that the appraisal of the mental characteristics of man is feasible. With many, indeed, the pendulum has swung so far in this direction that they are ready to accept as valid almost any kind of numerical finding based upon a "test." With naïve confidence they impute to the test result the same kind of significance that is suggested by the name its author has assigned to it, with little regard to the special factors affecting the individual instance. They overlook the fact that two tests bearing the same name may yield very diverse results. The sign is identified with the thing it is presumed to signify. Individual idiosyncrasies are lost sight of in the concern with mass tendencies.

This kind of wishful thinking on the part of the test users²⁴—educators, workers in the field of applied psychology, social case workers, vocational counselors—has arisen both from the urgent need these persons have felt for some means by which they might arrive at a better understanding of the persons with whom they work and from the discovery that the tests, at least to a certain extent, do fill that need. Moreover, the tests often appear to lend authority to pronouncements which otherwise might fail to carry conviction. Thus their popularity receives some further impetus from the effect of their use upon the self-esteem of the person using them. "The tests show . . ." is an argument few laymen will attempt to answer.

That this initial enthusiasm, even though it has led to some degree of erroneous thinking, has upon the whole been advantageous to the testing movement can hardly be doubted. No instrument will perfect itself while lying idle. The learner who receives too much adverse criticism for his inevitable blunders during the early stages of his learning is likely to become discouraged and quit. Later on, however, he can afford to take stock of his procedures and correct his errors. And in the field of testing we may safely say that the initial stages of learning by trial and error have now been passed and that it is time for us to enter more

²⁴ It is scarcely necessary to point out that there has been at all times since the beginning of the testing movement a large number of scientists who have labored zealously to analyze the sources of error in tests and testing procedures, to devise methods for their correction, and to point out the assumptions which underlie the various methods used and set limits to the interpretations which may be drawn from them.

earnestly upon the later phases in which we should review our progress, take note of our bad habits and erroneous concepts, and thereby bring a greater degree of insight to bear upon our methods. In the next chapter, therefore, we shall attempt a brief overview of the present state of our thinking and our accomplishments within this area, after which we shall turn to a consideration of some of the basic principles and theories underlying the construction of tests and their application to human beings of varying abilities and backgrounds.

The Present Status

RAPID GROWTH OF THE TESTING MOVEMENT

There is an old saying that if a man makes a better mousetrap than had previously been available, the world will make a trail to his door. This will be the case, however, only if the world has a crying need for mousetraps. If the need is sufficiently great and is not met in any other way, the mousetrap need not even be a very good one, at least as judged by more sophisticated standards. Moreover, when other potential makers of mousetraps see the procession of buyers arriving at their neighbor's doorstep, they will immediately set to work to try to improve on his model. It will not be long, one may be sure, before the market will become flooded with traps of all kinds, each claiming some point of superiority over all the others.

That is what has happened in the field of testing. Binet's tests, if judged by modern standards, were certainly not very good, but they filled the need for an objective means of appraising the abilities of children so much better than anything previously available that they were in immediate demand, and, as we have seen, attempts at improvement upon the model left by Binet were begun at once. The success of the newer tests of intelligence led, naturally enough, to the idea of devising methods for appraising other mental and behavioral characteristics. The need for some quick and easy method for classifying soldiers, which became apparent upon the outbreak of World War I, led to the development of group tests. The readiness with which the latter could be devised and standardized, at least after a fashion, encouraged many psychologists and educators to try their hand at test construction. All this has led to a multiplication of tests and testing devices that has hardly any parallel in the history of scientific method. Hildreth's *Bibliography of mental tests and rating scales* (Second Edition), which was published in 1939, has 4279 citations, not counting a six-page bibliography of bibliographies. Her *1945 Supplement* (Hildreth, 1946) raises this to a total of 5294 titles. Wang's list, published a year later (1940)

includes well over 5000 titles. South (1937) cited 5005 articles, most of them reports of new tests, which appeared during the fifteen-year period between 1921 and 1936. The *Review of Educational Research* publishes a triennial bibliography on tests and testing, which, as a rule, includes from 200 to 400 carefully selected titles. Although the spate of new tests which marked the period between 1920 and 1940 seems now to have passed the flood stage, it is probably a safe guess that even at the present time the average number published within a single year would run well over a hundred.

It is impossible to say how many psychologists and psychometrists in the United States are at present spending at least a part of their time in the administration, construction, or evaluation of tests. Among the first 500 names appearing in the 1948 *Directory of the American Psychological Association*, 252 either cite testing as one of their research activities or are personally known to me as actively engaged in such work. This figure certainly represents a gross underestimation of the number so occupied, inasmuch as the terms used for describing the kind of work done vary so greatly from one person to another. The 252 cases do not, for example, include those reporting "clinical work," "child development," "personnel counseling," and other terms which in most cases involve at least some work with tests and measurements, unless the latter are specifically mentioned. Of the 500 cases, 65 are listed by name only, and were presumably not actively engaged in psychological work at the time the *Directory* was published. These were entered on the negative side of the tabulation sheet, although it might perhaps have been better to omit them from the count. Moreover, there are a large number of persons who are not members of the Association but who are devoting all or part of their time to testing. It seems safe to say that well over half of American psychologists at the present time, as well as a good many persons who are not affiliated with any psychological organization, are doing some work in this field.

The forty-odd years which have elapsed since the appearance of Binet's first crude scale have thus witnessed a tremendous expansion of the idea and its applications. Not only tests of intelligence but those designed for the appraisal of almost every conceivable aspect of the abilities and behavioral tendencies of children and adults have been devised and more or less completely standardized and tested. The number of persons engaged in the construction and use of tests has shown an equally phenomenal increase. These testers and test makers vary in competence, and their products vary correspondingly in excellence. Although with the passage of time, increased competition has tended to weed out those of least value, with consequent steady improvement in the

general average, there are still far too many persons who enter the field of testing with inadequate preparation and who have little comprehension of the basic principles underlying the methods they use. It is highly unfortunate that the traditional standards for the training of mental examiners have been lower than those required for college teaching or other areas of psychological work. Inasmuch as such important decisions as whether or not a given child shall be sent to an institution for the feeble-minded or put in a special class, be sent to a reformatory or placed on parole, be given extra promotion or special tutoring, be regarded as a suitable or as an unsuitable candidate for adoption into a superior home, and other matters of equal weight for the individual often hinge upon the recommendations of the mental examiner, the need for training at least as thorough as that required for the prospective physician seems obvious.

SOME COMMON FALLACIES REGARDING TESTS AND TESTING

It was unfortunate that when, more than a quarter of a century ago, Terman first called attention to the fact that in most cases the intelligence quotients obtained by the use of the Stanford 1916 Revision with children of school age will not vary by more than a few points upon retesting with the same scale after intervals varying from a few days up to six or seven years, he used the expression "constancy of the IQ" to indicate this general tendency. Neither then nor at any subsequent time did he intend that this term should be interpreted in other than a relative sense. As a matter of fact, in his first formal presentation of his findings (Terman, 1919) he was careful to state the probabilities of a change of any given number of IQ points, as far as these probabilities could be determined by data then available to him. He pointed out that although about half of the children retested did not change their original standing by more than 5 points in either direction, about one out of six changed as much as 10 points and one of fifty (2 per cent) changed in apparent standing by 20 points or more. Although subsequent findings have shown the need for modifying a number of the conclusions reached by Terman at this time,¹ the surprising thing is that at a date when intelligence testing was still so new he should have been able to reach an estimate which agreed so closely, on the whole, with what has later been found. The unfortunate aspect is that his findings and conclusions have been so grossly misinterpreted by many well-meaning but poorly informed persons. The general tendency has been to exaggerate the extent of the

¹ See Chapter 11.

supposed "constancy" by forgetting all about the more variant half of the population reported by Terman and to assume that every IQ should remain "constant" within the limits of a 5-point range of variation. A second source of error comes from the tendency to think of an IQ as something resident in the child, not dependent upon the test used in deriving it. All sorts of tests are given, and the scores transmuted into "mental ages" and "IQ's," which are then expected to show no more variation upon retest than those derived from the Stanford. These assumptions are by no means confined to the ignorant; they are current even among college professors. Not long ago a well-known professor of education remarked to me that he had come to the conclusion that intelligence tests might better be given up completely because their results vary so greatly from time to time as to render them of little or no value for the guidance of children. As evidence he cited the "IQ's" obtained on successive tests by his twelve-year-old son since his entrance into junior high school. The range was from about 115 to over 150. Inquiry disclosed that with the exception of one Stanford-Binet given by a student in training, the so-called "IQ's" had all been derived from various group tests. The father had also failed to take account of the fact that in spite of the rather large numerical differences in the results, all the tests had agreed in placing the child well within the top 25 per cent of the general population.

It will be emphasized repeatedly throughout this book that while a score (however it may be expressed) on a test designed for the appraisal of some mental or behavioral trait may be, and often is, a useful indicator of the individual's standing, such an indicator is not necessarily identical with the thing it is presumed to indicate. Even the best of our present tests are subject both to *random errors*, resulting from crudeness in the measuring instruments and momentary shifts in the subject's attention and effort, and to *systematic errors*, the sources of which lie, for the most part, outside the immediate testing situation. Examples of the latter are the errors resulting from direct or indirect coaching of some subjects; unrecognized sensory defects which may lead to misunderstanding of instructions or to reading difficulties which hamper performance on group tests; marked differences in environmental background which make the items used in tests designed for the child of ordinary experience poor samples of the things which children with a different background of experience have learned to do; the use of poorly trained examiners;² and a host of other

² The poorly trained examiner usually makes both random and systematic errors in testing. The first result from fluctuations in procedure from time to time and in many cases from carelessness in scoring. Systematic errors arise from the bad habits which practically all such persons set up. Certain items will always be given or scored

factors. Test results cannot safely be interpreted by rule of thumb.

The appraisal of an individual may be based upon any or all of a number of different lines of evidence. At one extreme we have such varying and circumstantial sources as hearsay, casual observation of the individual himself or of his photograph, his handwriting or other personal productions, or records of his past achievements which are likely to be biased through the omission of some types of items and over-emphasis upon others. These are often given formal expression through the use of rating scales, questionnaires, and similar devices. At the other extreme we have the presumably (but not necessarily) more objective procedures known as tests, measurements, and "controlled" observations. The classification of methods in terms such as these, which depend upon their surface characteristics, is well known and need not long concern us here. But a more fundamental basis for classification in terms of what the user of a device conceives it to represent has rarely been stressed, and it is safe to say that few of the test makers or test givers have been cognizant of its importance. The question is: *Is the score obtained to be regarded as a sign, a sample, or a measurement of the characteristic named?*

Upon the answer to this question every aspect of test construction and interpretation depends. The selection of test items, the validation of the items and of the scale as a whole, the statistical handling of the results, and, most of all, the evaluation of their significance for an individual or for a group will vary according to the implicit assumptions upon which the test has been built up. Because these assumptions have commonly remained unrecognized, many errors of thinking have remained uncorrected, though their existence becomes apparent, once the basic hypotheses have been clearly formulated. In the following chapters, therefore, an attempt will be made to show how these differences in the nature of the tests affect their practical use and significance. It is hoped

incorrectly, and as a result the total score will be systematically raised or lowered. Moreover, the poor examiner is often inexpert in handling children and therefore may fail to secure their best interest and efforts. Some, however, go to the other extreme, and in their zeal to secure the best performance of which the child is presumed to be capable, give more than the permitted amount of help or urging.

Although the administration and scoring of group tests usually require less training than do individual tests, even here the poorly equipped examiner may adopt practices which have a systematic effect upon test results. More than once I have known an "analogies" test to be completely ruined by misplaced emphasis upon the words used as examples. It has been shown that praise or reproof can so modify effort in such a test as to bring about significant changes in the group averages (Hurlock, 1925). Thus the general attitude of the examiner, whether encouraging and cheerful or fretful and disheartening, may also be a more important factor in bringing about systematic differences in the results of group tests than many have supposed.

that in this way some of the prevalent misconceptions that have been our heritage from the past, when practical necessity too often led to the construction of devices that served an immediate purpose but lacked a firm basis in fundamental theory, may be cleared up and the way to future progress indicated.

PART II

Principles and Methods

The Bearing of Testing Theory upon Test Interpretation

DEFINING A UNIVERSE FROM SIGNS OR SAMPLES

A quarter of a century ago (Symposium, 1921), at a meeting held in Boston at which a number of the world's leading psychologists were invited to express their views as to the nature of intelligence, E. G. Boring created a mild sensation by proposing that the Gordian knot be cut once and for all by the adoption of a purely operational definition. In the physical sciences, so Boring pointed out, we gain our understanding of a quality by measuring it. Weight is the amount of gravitational pull upon a given mass which is registered on an instrument especially designed for measuring this pull. For most of us, therefore, weight is defined very simply. It is what the scales measure. In like manner, said Boring, we shall make sounder progress if we think of intelligence, or of any other characteristic for which we have designed at least roughly serviceable measuring instruments, in terms of *what the tests test*.

If Boring's proposal were to be adopted it would obviously necessitate a radically different approach to the selection of test items and to the interpretation of test results from that which has hitherto been customary. With the possible exception of the Rorschach Inkblot Test,¹

¹ Few, if any of the Rorschach workers seem to have understood clearly that the fundamental difference between this test and others lies in the fact that it deals with signs and not with samples. In tests of the more usual sort, the extent to which a given instrument really measures what it purports to measure is determined by comparing the scores earned by persons who are known or believed to differ with respect to the characteristic in question with their standing on other estimates or measures of the trait which the test presumably indicates. The meaning of the test is thus predesignated; the only question is: How well does it correspond to the criterion by which it is to be judged? The criterion is thus regarded as the independent variable, the test as the dependent variable. Or we may express the relationship in another way by saying that the criterion is the universe in which we are interested; the test is a sample of that universe. But if the test is not regarded as a sample but as a sign,

the meaning of a test score has usually been predesignated by the originator of the test. Not infrequently we have little more than a name, which is variously understood by different people, to inform us about the nature of the behavioral universe of which the test is presumed to be a sample. Even when the test maker has taken care to define his terms, the all too common practice of designating a test by some word in everyday use such as "intelligence," "egocentricity," "dominance," and the like, renders it altogether likely that many of those who make use of the method will interpret the results according to their own preconceived notions of the meaning of these terms which may or may not agree with those of the person devising the test.

Boring's suggestion implies that the test should be regarded as a sign and not as a sample. This means that the nature of the universe—that is, of the trait to be appraised—must remain undetermined until its character has been experimentally established. Instead of attempting to obtain representative samples of a universe with but vaguely defined characteristics and limits, one would begin with definitely known facts—the actual responses of a group of subjects to a series of relatively unstructured situations.² We should then look for common characteristics among those subjects whose responses have certain factors in common with each other. The meaning of these responses would then gradually

this relationship is reversed. The question then becomes: To what universe does this sign point? In the one case we start with a more or less clearly defined "trait" and seek for manifestations of it in the behavior of individuals; in the other case we begin with specified acts of behavior and try to find out their meaning in terms of a broader context.

It is unfortunate that many of the Rorschach workers have failed to see that their method necessitates the definition of a universe by means of the signs which point to it rather than the selection of a sample of a predesignated universe. For them the question is not: How do persons with these (presumed) characteristics interpret the inkblots shown to them? Rather it is: What kind of persons interpret the blots in this particular way?

² An "unstructured situation" is one which permits the subject to respond to it in a practically unlimited number of different ways, that is, one which leaves him free to create his own structure according to his own ideas. A test of the multiple-choice type, for example, does not permit the free expression of ideas on the part of the subject since the only alternatives open to him have been set by the test constructor, who also decides, as a rule, which of these alternatives is the "right" one, that is, most representative of the universe in which he is interested. However, a completely unstructured situation, such, for example, as is afforded by a lump of soft clay which is handed to the subject with no instruction whatever as to what he is to do with it, permits such an extremely wide range of possible responses that the task of finding out what they mean would be almost hopeless, and in all probability would not repay the effort since many of the responses would point to mere circumstantial factors rather than to stable trends in the individual personality. For this reason a partially or at least slightly structured situation, such as that provided by the Rorschach inkblots, is generally preferred to one in which no initial cues at all are provided.

emerge as, with the accumulation of data, hypotheses would be proposed and either verified or rejected. At the outset, while its meaning is still undetermined, the test would be given a noncommittal designation such as "Epsilon." "Epsilon" would then be defined in accordance with Boring's definition as "what the test tests." But as the test is given to more and more people, certain facts would, in all probability, begin to emerge. First it might appear that the responses of the same individuals to different items of the test do not follow a random order but tend to group themselves into classes. Accordingly, when it is found that Subject No. 24 gives Response A for Item 1, it can be predicted with better than chance probability that Item 2 will elicit Response B, and Item 3, Response C, and so on throughout the other items. But if, instead of responding to Item 1 with A, his response had been K, then the chances are that his responses to the other items would have been those characteristic of the group to which K belongs, with Item 2 eliciting Response L and Item 3, Response M. Certain types of responses would thus seem to belong together and therefore be regarded as different signs of the same behavioral universe.³

The universe to which the signs point may be discrete, that is, of an all-or-none character like an apple, a single chair, or an individual human being, or it may be a continuous attribute pertaining to all individuals of a class but in varying degrees such as beauty, intelligence, or mathematical ability. Until recent years mental testers have made few attempts at studying the signs which point to a discrete individual. Wolff (1943, 1946) and a few others have made a beginning in this direction, and other workers have made certain backhanded attempts at it by following a contrary procedure. Starting with an individual, they make a more or less elaborate analysis of his "traits" by finding his

³ The term "behavioral universe" is not identical in meaning with the word "trait," although the two expressions have a number of points in common with each other. However, since our concept of a "trait" is derived from casual observations, it is natural enough that we should include within it only those forms of behavior which bear some surface resemblances to each other, that is, those forms which we judge to have common characteristics or to stem from a common cause. Progress in the study of traits is therefore likely to be slow, since it is limited by human observation and is based upon rather shallow ideas of resemblances and differences.

In contrast, the characteristics and boundaries of a "behavioral universe" are determined empirically by the application of statistical procedures to a set of experimental data. The forms of behavior included within it may or may not appear to be similar; the only requirement is that the occurrence of one involves greater than chance likelihood that the others will also appear in response to the appropriate situations. Although progress in the study of these universes may be slow at first, the facts established in this way have the great advantage of verifiability and are also less likely to be influenced by conventional stereotypes of thinking and attitude.

relative standing on a number of different tests and measurements. Inasmuch as the meanings of these tests are in most cases only loosely defined, the "profile" method, as it is usually called, is less useful than it might theoretically become if more exact and uniform definitions could be ascribed to the "traits" upon which such profiles are based.⁴ The difficulty, of course, arises from the fact that practically all of our present-day tests have been constructed upon the principle of sampling rather than upon the principle of signs, and the same name has been given to universes that may and often do differ materially from each other.

THE PREDICTION OF BEHAVIOR FROM THE STUDY OF SIGNS

When the search begins at the other end, that is, with the sign, and the universe is defined as that to which the sign or the series of signs is found to point, the procedure differs in a number of ways from that followed in the more usual method of sampling. First and most obvious of these differences is the name likely to be employed in designating the universe thus identified. Instead of its being referred to as a "trait," which seems to imply something inherent in the individual, the chances are that it will be thought of in terms of its accuracy in predicting behavior. What can we infer from the sign and how dependable is the inference? Under what circumstances is the prediction warranted?

As far as the nature of the characteristic presumably appraised by a given test or scale is concerned, we may then note the following differences in the theories underlying the two ways of approaching the problem:

	<i>Method of samples</i>	<i>Method of signs</i>
Definition of universe	Predefined	Emergent
Limits of universe	Arbitrary	Empirical
Designation of universe	Usually an abstract noun	Behavioral in terms of probability
Interpretation of terminology	Varies to a greater or less degree with different workers	Comparatively uniform

THE PREDICTION OF BEHAVIOR FROM THE STUDY OF SAMPLES

According to Warren's *Dictionary of psychology* (1934), a trait is a

⁴ A further difficulty arises from the fact that as a rule the sampling of subjects used in obtaining the normative standards with which the individual is to be compared will not have been the same for all the tests used. Often the available information regarding the sampling is insufficient to enable one to judge whether or not the groups have been reasonably similar.

"distinctive mode of behavior of a more or less permanent nature, arising from the individual's native endowments as modified by his experience." Not all situations, however, give occasion for the display of any one specified mode of behavior, nor does a single episode provide sufficient evidence to permit an observer to make a valid judgment as to the likelihood of its recurrence.

In building a test by the sampling method the investigator is immediately faced with a dilemma. He must keep his procedure as brief as possible in order to avoid fatiguing his subjects or demanding more of their time than is likely to be available to him. On the other hand, he must secure enough data to constitute an adequate and representative sample of the mode of behavior in which he is interested. He cannot, therefore, afford any waste items; time is too precious. Because of this, the method of random sampling which will be described in the next chapter is obviously unsuited to his purpose. Too many of the items chosen in this way are likely to be of little value. He must therefore try to choose a series of items which will provide the best possible picture of the total universe in which he is interested. He must search for a representative sample.

Representative of what? It is surprising how few test makers have clearly faced this question. They have fumbled about with it by means of various statistical procedures designed to "measure" the test's "validity." Some have gone so far as to draw up formal definitions of the trait which they desire to appraise. Few, if any, have taken into account the elementary fact that neither the samples of behavior which they secure in the course of their testing nor the larger behavioral universes of which such tests are presumed to be samples can be usefully considered without reference to time and circumstance. Are these universes static or changing? If changes occur, what is their rate of progress? Are the changes uniform for all subjects? To what extent are they determined by forces exerted by external agents of an unknown or unpredictable character and to what extent by factors inherent in the organization and structure of the universes in which they take place?

These are by no means idle questions. Consider, for a moment, how they apply to the question of intelligence testing. There can be no doubt that the forms of behavior which we regard as "intelligent" change in a reasonably consistent manner as age advances. Indeed it is because of these systematic changes in behavior as the individual grows older that the concept of mental age was advanced. There is likewise no doubt that such changes take place most rapidly during the early years. If, then, the nature and extent of these changes are determined chiefly or wholly by

inherent factors, a sample of the behavior taken at one age may be a sufficiently good sign⁵ of that likely to be shown at some later age to form a reasonably accurate basis for prediction. But if, on the contrary, the changes result for the most part from external causes that cannot be predicted in advance of their occurrence, then, no matter how representative and adequate a sample of the forms of behavior shown in infancy may be secured, their significance will be only temporary. They cannot be regarded as signs of the kind of behavior which will appear at a later age.

However we must not make the mistake of assuming that absence of correlation between samples of behavior taken at earlier and later ages necessarily means that external factors have been responsible for the changes. The samples may have been drawn from different behavioral universes and not, as we have naïvely assumed, from the same universe sampled at different stages of its development. In spite of repeated cautions the "naming fallacy" continues to obstruct scientific progress in many areas, and in none, perhaps, more glaringly than in the field of mental appraisal. Two tests are called "intelligence tests." It is accordingly taken for granted that they sample the same universe. A child of two is "tested," that is, a sample of his abilities at this age is obtained. At the age of ten he is retested by means of a test called by the same name but for obvious reasons made up of an entirely different series of tasks. The assumption that each is an equally representative sample of the *same* universe is, if we examine the facts, grounded more upon the

⁵ In a sense, of course, any sample may be regarded as a sign of the universe from which it is drawn, inasmuch as we make inferences about the total on the basis of the portion of it which we take as a sample. We may say, then, that any sample may also be looked upon as a sign, but there are signs which point to a universe of which they are not a part. It can be shown, for example, that from a man's occupation one can predict with better than chance success what will be the intellectual level of his children. From a knowledge of a man's height we can make an estimate of his most probable weight. Further information will enable us to determine the amount by which our estimate is likely to be in error.

The basic difference between the two, however, is not a matter of terminology but of method. It is a question of the point at which an investigation starts and the direction in which it moves. In the case of sampling we begin with a universe about which something is known, surmised, or hypothesized, and attempt to analyze it into its component parts. We then select a proportional number of samples from each part to make up our test. The success of the procedure obviously hinges upon the accuracy of our information regarding the universe and of the parts which constitute it. In the method of signs we begin with known facts and look for relationships between them. Gradually, as patterns or clusters of related items are discovered, the universe to which they belong or toward which they point will become manifest. The former is an analytic method; the latter a synthetic one.

faith inspired by the use of a common name for the two series of tasks than upon objective evidence. That some overlapping of the two universes really exists is indicated by the positive correlations usually obtained at the ages specified. That the two are by no means identical is manifested by the relatively low magnitude of these correlations, which have commonly been found to be much smaller than that to be expected on the basis of randomly distributed errors of sampling. For example, Bayley (1940), in reporting the results obtained from a nine-year study⁶ of the development of a group of sixty-one children from birth onward, found a zero or a slightly negative correlation between scores earned on the California First Year Mental Scale before the age of one year and those obtained on the Stanford-Binet at four years or older. That is, the behavior samples obtained for the infants could not be regarded as signs from which those to be obtained at a later age might be predicted with better than chance success. There are three possible explanations for this. Intelligence as manifested in behavior may not, during its early stages, conform to lawful and predictable rules but may vary in its growth in an erratic and undeterminable manner. While this explanation is possible, it seems decidedly improbable in view of the consistency and regularity of other developmental processes not only during infancy but even during prenatal life. A second explanation is that intelligent behavior is wholly a resultant of the kind of external stimuli the child receives. This is a highly controversial issue which will be discussed in some detail in a later chapter. For the moment we shall only note that such an explanation is hardly consistent with the fact that test samples obtained at a slightly later age, say after the age of two years, do enable us to predict within broad limits that which will be manifested later on. Samples taken at four years or later commonly afford sufficiently dependable signs of those to be obtained later, even after a lapse of ten years or more, to serve many practical and scientific purposes (Bayley, 1940; Bradway, 1944; Ebert, 1941; Goodenough and Maurer, 1942). It is not easy to see why such environmental differences as exist within the range of what would ordinarily be regarded as "good" homes should be all-determining during the first year and relatively unimportant after the age of four. Logically, perhaps, one might expect that such factors as the number and quality of books in the home, the willingness and ability of parents to answer the child's questions, and the effort made to stimulate his intellectual curiosity would have a more potent influence upon the pattern of mental growth after the age of four years than

⁶ This investigation is still in progress.

before, but this does not appear to be the case if the statistical evidence is to be trusted.

A third and perhaps more probable explanation is to be found in the tests used at the different ages. As was noted before, the assumption that the behavioral universes sampled at these ages differ from each other only in respect to the stage of development that has been attained is based upon very insecure evidence. For the most part, this evidence consists of some degree of superficial resemblance between the tasks used, or the opinion of the test constructor that a given bit of behavior is a sign of "intelligence," together with such ambiguously interpretable facts as the following: different babies vary in their ability to perform (or at least in their performance of) these tasks; some internal consistency can usually be noted in the performance of a given child at a stated occasion on a series of tasks of this kind as well as some similarity between his performances on two or more occasions separated from each other by only a brief interval of time. Motivational factors, however, may be largely responsible for the two facts last mentioned. That superficial resemblance between two tasks is no guarantee of their essential similarity has been cogently pointed out by Bayley (1940) as follows:

It seems obvious in the block-building series, for example, that the initial stages, which require insight into the processes of putting one block on top of another and then letting go of it while it is still there, are very different from the muscular control and persistence required for stacking eight blocks on top of one another; and these in turn are different from the ability to see the relation of the separate blocks to the whole structure when the child tries to copy the five-block "gate" or to reproduce the stairs from memory. Perhaps there are similar but less obvious differences in the functions measured by the tests of vocabulary at different ages. The question here is whether it is in the very nature of intelligence that these functions differ at different ages, or whether similar functions are present at all ages—though still not isolated in such a way that they can be compared.

THE QUESTION OF TRAIT NAMES

One can hardly emphasize too strongly that every test constructed on the theory of sampling must be appraised on the basis of its conformity to the universe of which its author designed it to be a sample, and that only the author is qualified to describe and delimit this universe. It not infrequently happens that when the author has designated the universe which he desires to sample by some term in everyday use, others may not agree with his terminology. For example, the author who devises what he calls a test of "introversion-extroversion," or perhaps of

"dominance-submission," or "emotional stability," may have a very different concept of the meaning of these terms from that held by another person who makes use of his test. Two different authors, each of whom devises a test called by one or another of these names, may nevertheless differ so greatly in their ideas of the universes to which the name applies that the sample tasks which they include in their tests may have little or nothing in common with each other. Thus, even though the items included in each test yield results that are internally consistent, and though retests of the same subjects by means of either test classify them in much the same way on both occasions, there may be little or no resemblance between the scores obtained when the two tests are compared with each other. Merely calling two things by the same name does not make them equivalent or even similar.⁷

The person who insists upon substituting his own concept of what a given behavioral universe *should* be like for that held by the person who has devised a test according to a different understanding of the term by which the test is designated is therefore doomed to disappointment from the start if he attempts to make such a test serve his purposes. No test should be condemned because it does not chance to conform to some special pattern that the prospective user has in mind. He is under no compulsion to use it if it does not meet his requirements, but he has only himself to blame if he fails to ascertain in advance, as nearly as available data permit, what are the characteristics of the universe which the test was designed to sample. Unfortunately, too many test makers have themselves been hazy in respect to this all-important point and have failed to supply the information needed for a clear understanding of the test which they have been at such pains to develop. They overlook the fact that when it comes to a matter of trait names, a given language is far from being the common possession that its users generally suppose it to be.⁸

THE EXPERIENTIAL REFERENCE

In the early days of intelligence testing, but slight attention was given to the fact that, as implied in Warren's definition cited on page 101, experience as well as inborn tendencies plays a part in molding the

⁷ One of the stories attributed to Abraham Lincoln is appropriate here. In the course of a friendly argument, Lincoln asked his opponent, "Now tell me, if you call a tail a leg, how many legs has a dog?" "Five," was the ready answer. "Not a bit of it," replied Lincoln. "You can *call* a tail a leg as much as you please but that won't make it any more use for walking on."

⁸ A very complete list of trait names has been prepared by Allport and Odbert (1936).

behavior of an individual. This is equivalent to saying that it is unsafe to apply the same criteria to the appraisal of individuals or groups of radically different background. In the world of everyday affairs, some recognition of this fact usually is taken. We say, "You cannot expect that child to have good manners; he has never been properly trained." Or we note that of two boys who are unable to read, one has been reared in a backward community and has never had a chance to attend school while the other has failed to learn in spite of every opportunity to do so. The objective facts are the same for both; a test of reading achievement would yield similar results in each case. But the significance of these facts is surely not the same.

In describing the universe of which a given test is presumed to be a sample, it is therefore necessary to take account not only of the *kind of behavior* to be included but also of the *kind of persons* who may be expected to behave in this way. It is doubtful whether any test has ever been devised that can fairly be said to have the same significance for all persons. It therefore becomes incumbent upon all test makers to state, as exactly as possible, the limits of the group for whom the test is appropriate. The behavior and the behaving individuals are by no means independent of each other but must be considered simultaneously. For example, Terman (Symposium, 1921) has defined intelligence as "the ability to think in abstract terms." We must therefore assume that this is the universe which he has intended to sample by means of his tests. But since thought cannot be observed directly but can only be inferred from the overt behavior which is its end product, and since abstractions are the derivatives of concrete experience, they are likely to differ in form according to the nature of these experiences. It is evident, then, that a stated degree of ability to deal with a given set of test items is not likely to have uniform significance unless the experiential background of the subjects is reasonably similar. In terms of sampling theory, we may note that the group of subjects upon whom a given test is standardized must be regarded as a random sample of the universe of persons to whom the test is known to apply. The character of the universe is determined by that of the standardization group. It is, of course, true that the same condition holds with respect to the behavioral universe of which the test items are presumed to be samples, for no matter what concept of that universe the test maker may have formed, if his selection of samples is biased, the judgments derived from them will be biased in a similar manner. This point, however, is better understood; it is inherent in the familiar distinction between the "reliability" of a test and its "validity."⁹

⁹ Otis defines "reliability" as "the consistency with which a test measures whatever it does measure" while "validity" is said to mean "the degree to which a test measures that which it purports to measure."

But the fact that a given test may be both "reliable" and "valid" when used with a group of subjects for whom it is appropriate, and neither "reliable" nor "valid" when applied to a group of different composition, is frequently overlooked.

When a test is given to persons who were not a part of the original standardization group we must therefore be able to assume that such persons are sufficiently similar to that group in respect to experience to justify our assumption that they form a part of the same universe.¹⁰ For whether we are aware of it or not, such an assumption is implicit in every application of a given test to an individual or to a group. One of the unfortunate results of the common practice of designating tests in terms of some abstract quality such as "intelligence," "dominance," and the like, is the tendency to reify¹¹ such terms into some kind of human attribute unaffected by circumstances. Accordingly, tests designed for use with a certain class of subjects (as, for example, English-speaking urban school children) have frequently been applied to groups of such widely different background from that of the original group that to regard them as unbiased samples of the same universe (that is, to interpret the results according to the same standards) is unwarranted.

If the problems of sampling constitute such serious hazards in the work of test construction, it is natural to inquire why this method has been so generally adopted. The answer is to be found in the fact emphasized in the first two chapters of this book. Tests were not originally constructed or conceived as instruments of precision to be used in the answering of scientific problems. They were designed as tools to serve certain immediate and very pressing needs. It was these needs that dominated the minds of the earlier test makers; it was but natural that a concept, rough and ready though it might be, of the universe which it was *desired to appraise* should take precedence over the recognition and study of signs which might point to some other universe of no immediate interest. Practical necessity rather than scientific curiosity as such was the dominant motive in the construction of the early tests. Moreover, the search for signs as exemplified by the phrenologists, the anatomists, and the psychophysicists of the latter part of the nine-

¹⁰ It may be well to point out once more that I am not here referring to the controversial issue of whether or not the ability to think in abstract terms is affected in a quantitative as well as in a qualitative manner by differences in experience. That question will be discussed in a later chapter. At the moment we are concerned only with the influence of experience upon the *kind* of material with which the thinking individual deals most readily.

¹¹ *Reification* means ascribing the qualities of reality to that which has no real or independent existence. For example, we *reify* intelligence when we think of it as "something" possessed by man instead of as a term used in describing his behavior.

teenth century, who hoped to find in their measurements a way of predicting scholastic success or other indications of the ability to think in abstract terms, to reason, and to form sound judgments in areas outside the field directly studied, had been so uniformly unsuccessful as to discourage further effort along these lines. When Binet demonstrated that the method of sampling worked, at least reasonably well, his discovery was hailed with acclaim, since the universe which he succeeded in sampling was one of great practical import and was well enough defined to serve the immediate needs of those who used his methods. As time passed and enthusiasm mounted, other people with different concepts of this universe added their sets of sample tests to those already on the market. Statistical procedures multiplied. The gradual acceptance of the idea that in the proper application of statistical method to the data obtained by tests was to be found the key by which the secrets of human mentality might eventually be laid bare unquestionably led to great improvements in the techniques of measurement. And there is no doubt that the use of statistical methods in the field of testing has enabled us to find at least tentative answers to many questions of scientific and practical import, the scope of which extends far beyond the early concept of testing as an aid to the understanding and guidance of the individual. The mental test, as Terman (1924) hoped, has truly become a scientific instrument.

Its use, however, is not confined to scientists. Testing procedures and statistical methods of handling the results are learned, parrot-fashion, by many who have little comprehension of the limitations and pitfalls of the procedures they use so glibly. As a result, many erroneous conclusions have crept into the literature, a number of which occur and recur in varying connections and in slightly altered forms as the same stereotyped procedures are followed without regard to their suitability, and as interpretations are made without reference to limiting factors arising from the nature of the particular sets of data from which the conclusions are drawn. Because most of our present methods of testing are essentially sampling procedures, an understanding of the basic assumptions of sampling is of prime importance both for the constructor and for the user of tests. We shall accordingly devote the following chapter to a consideration of the principles of sampling in relation to the special field of tests and measurements.

Problems of Sampling

THE TEST AS A SAMPLE OF ABILITIES OR CONDUCT

As was pointed out in the last chapter, the question of sampling in mental testing has a twofold aspect. First, there is the matter of test content, since the items of which a test is made up are presumed to be representative of the total universe or "trait" which the author designs it to appraise. Second, there is the question of the kind of persons chosen to provide a standard sample of responses which are to constitute the "norms" with which the responses of other individuals are to be compared. Because these problems are so intimately related to each other we shall begin our discussion by pointing out certain general principles that apply to all questions in which sampling is involved.

The purpose of sampling is to secure information about the characteristics or quality of a total, too large to be conveniently examined in its entirety, through the inspection of some part of it which is presumed to be representative of the whole. The grain dealer does not examine the entire contents of a load of wheat offered for sale. Instead, he inserts a measure into each sack and removes a specimen which is assumed to be a fair sample of its entire content. If the consignment is large he may content himself by taking only a small number of sample sacks for examination. As a rule he will be reasonably safe in assuming that those portions of the grain he does not take time to inspect are similar in quality to those which he does examine. In like manner the housewife does not examine in detail each individual bean among the five pounds which she proposes to buy. She merely removes a handful from the grocer's barrel and judges the quality of the total on the basis of this sample. The method of sampling is much used in economic and sociological studies as well as in many aspects of daily life. Provided that sufficient precautions are taken to guard against bias, that is, to make sure that the sample really does resemble the universe of which it is a part closely

enough to meet the needs of the particular occasion, the method is not only convenient but dependable.

But universes differ in many ways that affect the difficulty of securing unbiased samples of them. In this connection the following points are to be particularly noted.

1. A universe may be of finite or infinite size. The living population of a given city on a stated date is an example of the former; the number of possible throws of a die¹ is an example of the latter. In the case of a finite universe, the usual reason for making use of samples is economy of time and effort. If the universe is infinite, a sample is all that can be had.

2. A universe may be homogeneous in composition and structure or it may be heterogeneous to a greater or less degree. If heterogeneous, the various parts or aspects may be evenly distributed throughout the total or they may tend to cluster in groups or strata. If the universe is homogeneous like chemically pure water kept at a constant temperature, a very small sample will suffice to give an adequate picture of the total. If it is heterogeneous like sea water in which various minerals have been dissolved, greater care is needed to avoid bias in sampling, especially if all the parts are not kept in a constant state of mixture. If not thoroughly mixed, the concentration of minerals is likely to become greater in certain regions than in others. Among human beings, the tendency for people of like characteristics to congregate in groups makes it far from easy to secure a truly representative sample of any given universe because of the ever-present likelihood that too large a proportion of the sample cases will be drawn from one center or stratum, with consequent underrepresentation of others. Such a sample is said to be *biased*. The likelihood of bias is so great, especially where human characteristics are concerned, that a great deal of careful investigation has been devoted to finding methods by which such bias may be minimized or avoided when the aim is to secure a truly representative sample of a given universe, or to determining its presence when the interest centers about the peculiarities of a given sample rather than about the universe as a whole. In general it may be said that the more heterogeneous the universe becomes and the less uniformly distributed are its parts or elements, the more difficult it is to secure an unbiased sampling of it and the greater, consequently, must be the number of separate samples needed to secure a truly representative picture of the total.

3. Much or little may be known in advance about the characteristics

¹ In a sense, of course, this is not a truly infinite universe, for in time the die would wear out. If, however, we think of hypothetical throws rather than of tossing a material object, the number becomes truly infinite.

of the universe. The greater the amount of previous knowledge, the less danger there is of bias in sampling, provided that this knowledge is intelligently applied in deciding upon the method of securing the samples. Suppose, for example, that the problem to be studied is the average income of the families in a certain city and that it is tentatively decided to take as a sample the residents of every tenth house. If it so happens that a majority of the city blocks include exactly ten houses, such a method of sampling would obviously include too large a proportion of corner houses,² which are likely to be inhabited by the higher-income groups. If this fact were known in advance, some other method of selecting the sample could be chosen. It is because too little is definitely known or has been generally agreed upon with respect to the kinds of behavior classified under such general heads as intelligence, leadership, aggressiveness and the like, that the problem of obtaining representative samples of them in the test situation becomes so difficult.

4. A universe must possess a certain internal coherence in order to be classed as such.³ This coherence may be derived from setting boundaries, as when a pasture is defined by saying that it includes all the land enclosed by a given fence, or, less precisely, in terms of that which tends to cluster about a given center or nucleus. In dealing with material universes we are in most cases able to couch our definitions in terms of boundaries expressed in units of time and space. In dealing with behavioral universes we are more likely to resort to terms that imply some kind of common characteristic or focal point about which behavior episodes cluster. As a rule we are unable to set precise boundaries for the forms of behavior to which trait names are applied. We cannot say, "Up to this exact point the universe which we call *intelligence* runs, but no further." The boundaries of a state or a city are defined by their officially recorded limits; the boundaries of a specified day are set at the hours of midnight according to a standard timepiece. But how shall we set the boundaries of leadership? We cannot say. Our only way of defining a universe of this kind is by means of describing its most salient features, the nodal points to which actual behavior episodes conform more or less closely. By means of certain statistical procedures⁴ it is

² Or too small, depending upon the point at which the counting is started.

³ The word "universe" is derived from *uno* (one) and *vertere* (to turn or to turn into); hence that which has turned into or become one, a totality.

⁴ Such, for example, as the various methods of studying internal consistency by the comparison of scores on a single item or section of a test with each of the others separately or with the total of all the others, factor analysis, analysis of variance and covariance, etc. For an account of these procedures the reader is referred to any recent textbook on statistical method.

possible to "purify" a universe to a certain extent, that is, to make it more nearly homogeneous,⁵ but the complicated nature of the relationship between the character of an act and the characteristics of the person who performs it renders it unlikely that any one method can be devised which will permit as precise a definition of the limits of any particular ability or behavioral tendency as those which are available in the world of concrete affairs.

The misunderstandings and disagreements with respect to tests and their interpretation thus stem from a number of causes. In the first place, the universe of which a given test is presumed to be a sample is infinite in size since it includes all possible acts in which the alleged trait is manifested.⁶ Moreover, few if any of the traits for which tests have been devised are truly homogeneous in character, nor are their manifestations evenly distributed over all the acts performed by a given individual. A further source of disagreement and confusion is our ignorance of many if not most of the facts that should be known about these universes if bias in sampling is to be avoided. Finally, because their boundaries are indeterminate, some disagreement as to the areas which should be included within each universe is well-nigh inevitable.

When all these and other possible sources of error in sampling are taken into account, what is surprising is not that tests sometimes yield results that do not accord with other known facts concerning the individuals who are tested or that people often disagree about their interpretation or applications; instead, it is the fact that tests have proved to be among the most useful instruments ever devised for the study of human behavior. Bias in sampling is more than a possibility. It unquestionably appears both in the selection of test items and in the uncritical application of testing procedures to individuals or to groups for whom these measures are unsuitable. A test may be a poor sample of the abilities of a particular individual even though it is well suited to the majority. In many of these cases, intelligent examination of the known facts concerning the test and the person to be tested would indicate this fact in advance, and if this were always done many misunderstandings

⁵ Probably the best example of such an attempt at "purification" of a test that has appeared to date is Thurstone's (1938, 1946) work on primary mental abilities described in Chapter 15.

⁶ For the sake of brevity I have frequently used the word "trait" as if it were something really existent within the mind or muscles of man. The reader should constantly remind himself that a trait is nothing more than a group of actions which, because they have certain features in common with each other or are believed to stem from a common cause, are looked upon as constituting a behavioral universe. Although traits may and frequently do overlap each other, each presumably has certain unique features of structure and organization which distinguish it from all the rest.

and unsound interpretations could be avoided. No one should be judged by the standards of a group to which he does not belong.

METHODS OF SAMPLING: I. RANDOM SAMPLING

Two methods of obtaining a fair sample of a finite universe are in common use. These are known respectively as *random sampling* and *stratified* or *representative sampling*. In random sampling the essential requirement is that every item in the total shall have an equal chance of being included in the sample. Random sampling is commonly used when the universe to be sampled is relatively homogeneous throughout, or when the various parts of which it is composed are evenly distributed within the total without tendency to cluster in groups. An additional requirement is that all parts shall be equally easy of access. These conditions rarely exist in the case of human beings unless the population to be sampled is completely known. In such cases a model universe can be made up by means of a card file in which a separate card is used for each member of the universe. By means of thorough shuffling of the cards, such groupings as may exist within the actual universe may be broken up and distributed at random throughout the series. A selection, let us say, of every tenth card⁷ would then meet the necessary conditions for randomness. Such a procedure is useful, for example, in taking a sample of a college population where definite bias would be likely to result if the sample were drawn from the members of a single class or department but access can be had to the entire group. A random sample of adequate size drawn from a well-shuffled file of the entire college population can generally be depended upon to yield the desired information. The degree of precision of the estimate can be determined statistically, provided that there has been no bias in the shuffling, or it may be tested empirically by first returning the sample already drawn to the file, reshuffling, and then selecting a new sample to be compared with the first one. By repeating this procedure a number of times, the amount of variation from sample to sample can be ascertained. It is important that each sample be returned to the file before selecting a new one; otherwise the universe from which the second sample is drawn would not be the same as that from which the first was taken.

Another method of securing a random sample, which is sometimes

⁷ The proportion of the total needed to constitute an adequate sample will depend, generally speaking, upon three factors: (1) the degree of accuracy required, (2) the amount of variability within the universe, and (3) whether or not the sample is intended to be a miniature representation of the universe in respect to variability as well as in respect to the average or most typical of its qualities. If this is the case, a larger sample will be needed.

preferred, consists in the use of a table of random numbers.⁸ Again it is necessary to have a list of all the items included in the universe. The cases are arranged in any convenient order and numbered consecutively. By consulting the table, a sample is selected of which the numbers have been arranged to constitute a randomly occurring series. Again the extent of the sampling error may be determined either statistically or empirically by comparing different samples with each other.

METHODS OF SAMPLING: II. STRATIFIED SAMPLING

In many cases, however, it is either impossible or very inconvenient to secure an advance list of all the items in a universe to be studied. In dealing with human subjects whose tendency to cluster in groups, of which the structure is determined both by geographical and by psychological and sociological factors, is too well known to require elaboration, it is particularly unsafe to try to secure a random sample from such a stratified universe. In the examples just given, the universes were randomized artificially through the shuffling of cards or the use of random numbers. In working directly from actual cases where the original clusters cannot be broken up by such means, the method of stratified sampling is generally to be preferred.

In stratified sampling, the fact that the universe to be studied is made up of various subgroups or clusters that differ from each other in ways that may have a bearing upon the topic of investigation is recognized and dealt with directly. It is known, for example, that different occupations make differing demands upon abstract intelligence. As a result, men of differing intellectual levels tend to find their way into those occupations for which they are intellectually suited. Since children are likely to resemble their parents in mental as well as in physical traits, paternal occupation may be used as a criterion for dividing children into intellectual strata for purposes of sampling. Of course the differences between groups classified in this way are less marked among the children than among their fathers, but they are far greater than most people suppose, at least as far as the scores earned on standard tests of intelligence are concerned. The differences between the average IQ's of children whose fathers belong to the learned professions and those of the children of day laborers have commonly been found to run as high as 20 to 30 IQ points (Haggerty and Nash, 1924; Goodenough, 1929; Terman and Merrill, 1937; and others). Inasmuch as these groups tend to be separated geographically as well as psychologically through residing in

⁸ Tippet's (1937) lists of random numbers are most often used.

different parts of the city, attending different schools, and so on, it is easy to see that unless much care is taken to ensure that all are represented in their correct proportions, considerable distortion of the standards presented for any test in which abstract intelligence is a factor can easily occur.⁹

The method of stratified sampling consists in breaking up the universe to be sampled into the various layers or clusters of which it is composed, ascertaining what proportion of the total is included within each of these subgroups,¹⁰ and then selecting a sufficient number of cases to keep the proportions within the sample the same as those in the universe as a whole. Further stratification within the subgroups is often desirable. For example, suppose it is desired to secure a representative sampling of the adult working population of a particular city. From the census data it is ascertained that this population is made up of 80 per cent Whites and 20 per cent Negroes. The trait to be studied may reasonably be expected to vary not only with respect to race but also with sex and social status. It is decided to use the Taussig Industrial Classification as the best available indicator of the social level.

The most nearly independent of the three factors seems to be race, so the first stratification will be made on that basis. Sex will be considered next, since the proportions of the sexes who are employed outside the home may not be the same for Negroes as it is for Whites. In fact, the census data show this to be the case. Since the distribution of occupations is likely to vary both with race and with sex, this factor will be considered last.

The first step is to draw up a design for the stratification in terms of the proportions given by the census for the entire population of the city.

⁹ The distribution of occupations for employed men either in the United States as a whole or in any of its major divisions can be determined from the census reports. The census data have been used as a basis for controlling the sampling in the standardization of at least two modern intelligence tests—the Minnesota Preschool Scales in which the employed male population of Minneapolis and St. Paul was the universe to which the sample was made to conform; and the 1937 revision of the Stanford-Binet, where the entire population of the country, both urban and rural, was taken as the universe.

¹⁰ In deciding what subgroups are to be considered, two factors must be taken into account: (a) the pertinence of the classification with reference to the problem to be investigated, and (b) the availability of information as to the proportions of these groups within the total universe. The first point hinges upon the investigator's acquaintance with his problem. The second often appears more difficult than it is. Those not thoroughly acquainted with the census reports are likely to be surprised at the wide range of specialized information to be found both in the main volumes and in the various special studies published from time to time. Government bureaus, both federal and state, publish a good many bulletins that are useful for such purposes. For studies having only a local reference, the local chamber of commerce is often able to supply needed information or to indicate where it may be obtained.

The occupations listed in the census must first be reclassified in terms of the five groups proposed by Taussig (1920).¹¹ Although questions may occasionally arise as to the proper class in which a given occupation should be placed, the grouping can nevertheless be made with sufficient accuracy to serve most purposes. The complete design, expressed in percentages of the total population of the city, might appear something like the following:

Taussig Class	White (80 per cent)		Negro (20 per cent)	
	Male (60)	Female (20)	Male (12)	Female (8)
I	20	4	6	4
II	20	4	2	2
III	10	2	2	1
IV	5	5	1	0.5
V	5	5	1	0.5

A representative sample would then be made up in such a manner as to keep the proportions within each class the same as those indicated by the design. Let us assume that this sample is to consist of 500 cases. A table should be constructed, showing the actual numbers in each group.

Taussig Class	White		Negro	
	Male	Female	Male	Female
I	100	20	30	20
II	100	20	10	10
III	50	10	10	5
IV	25	25	5	2 or 3
V	25	25	5	3 or 2

Theoretically, each of the final subclasses within the sample should be chosen at random from the total representation of that class within the universe as a whole. In practice it is usually necessary to depend somewhat upon individual judgment in order to avoid the inclusion of

¹¹ Taussig refers to these as "non-competing" groups because of their relative independence and the fact that there is normally only a small amount of shifting from one class to another, although considerable movement may take place within each class. Roughly the five groups may be designated as follows: I. Day laborers; II. Unskilled workmen; III. Skilled workmen; IV. Lower middle class including clerical and semi-intellectual occupations; V. Well-to-do including professional and business men and managers of industry. Taussig's descriptions of each class, which are fairly exact and detailed, should always be consulted before attempting to make practical use of the system.

persons who are clearly not typical members of the class to which they chance to belong, particularly when the total representation of the class in question is small. As far as extreme cases are concerned, this is usually not as difficult as it may seem. A college graduate whom unusual circumstances have forced into a pick-and-shovel job is certainly not a typical example of Taussig's Group I. Unless the sample is so large that the law of probabilities would justify the inclusion of a few such exceptional cases, it is usually better to omit them and substitute others better suited to represent the class.

THE STANDARD SAMPLE AS A SOURCE OF REFERENCE

The numerical score obtained on a test does not become practically meaningful until it has been transmuted into some kind of interpretative expression by means of which the individual's standing within the universe of which he is presumed to be a part can be determined. To most people the statement that the height of a boy of seven years is 46 inches will not mean very much, but when they are told that 75 per cent of the boys of his age are taller than he is, the figure at once takes on significance. But before such a comparison can be made it is necessary to have learned the heights of each member of a representative sample of seven-year-old boys. The selection of a standard sample for the establishment of normative values with which the standing of others is to be compared is a task of the greatest nicety which demands far greater care than it has frequently received. In the past the emphasis has generally been placed upon the size of the normative group with but slight attention to the source of supply. Had it not been for the fortunate fact that public schools are likely to be the most convenient places for securing data on school children, colleges for securing information about college students, and so on, discrepancies resulting from bias in the reference samples would unquestionably have been far greater than they have usually proved to be. Nevertheless, even public schools and colleges differ sufficiently, one from another, to introduce a disturbing amount of bias into any sample that has not been carefully chosen to ensure its similarity to the universe which it is intended to represent. Maller (1933), for example, showed that when the IQ's of all the fifth-grade children in New York City were averaged according to the district in which their schools were located, the means for different parts of the city ranged from 74 to 118. Sangren (1929) found that when 100 seven-year-old children were given seven primary intelligence tests within a three-day interval, the average "mental age" of the group, according to the pub-

lished standards for the different tests, varied by as much as a full year, or 14 per cent of the chronological age. Inasmuch as such expressions as "mental age," "IQ," "standard score," and so on, are not absolute values but are derived by comparison with the performance of the standard samples used for obtaining norms, it is evident that the meaning of these terms cannot be uniform if the standards are unequally biased.

A striking illustration of the fact that the size of the sample is no guarantee that it will provide an unbiased picture of a given universe was afforded by the *Literary Digest's* straw vote on the presidential election of 1936. Although more than a million votes were secured, the results of the actual election were the opposite of what had been indicated by the *Digest* poll. In this case the failure of the sample to conform to the universe arose from the fact that the list of persons invited to participate was made up largely from telephone directories and similar sources in which the lower-income groups were poorly represented. Inasmuch as on that occasion the direction of voting was associated with the level of income, the size of the sample could not make up for the social bias. A much smaller sample secured by Gallup of the Office of Public Opinion Research according to the principle of stratified sampling predicted the results of the same election with remarkable success. However, the results of the 1948 election provide an equally striking demonstration of the fact that even carefully laid out plans for securing a stratified sample are unlikely to work unless a sufficient amount of information about the universe to be sampled is available. Certainly the populations who were questioned in that year by *Fortune Magazine*, the Office of Public Opinion Research, and other agencies of a similar type did not correspond to that which actually voted on Election Day. There are at least three possible reasons for the discrepancy between predictions and outcome. In the first place, inasmuch as the number of persons actually questioned was small in proportion to the number of voters, chance variations of sampling may have been responsible. While this explanation is possible, it appears unlikely in view of the general agreement in the results obtained by the major polling agencies. It is also possible that the information possessed by the agencies concerning the group of qualified voters was either inadequate or incorrect. This would lead to the inclusion of incorrect proportions within the various subgroups¹² represented in the samples questioned. A third possibility has to do with unforeseen shifts in the proportions of the various classes of qualified voters who exercised their right to vote.

¹² The number of factors possibly related to voting tendencies is so great that the choice of the proper subgroups is not easy. Age, sex, place of residence, income level, and usual political affiliation are among those customarily considered in the prediction of election results.

Finally, the explanation may be found in possible bias in the selection of persons to be interviewed on the part of those who conducted the interviews. According to the method used by most of the polling agencies who have relied on the method of stratified sampling, those who secure the data choose their subjects on the basis of their conformity to certain general characteristics laid down by the central agency. For example, a particular worker may be told to question six males between the ages of twenty-five and forty, all of whom are factory employees earning \$40 to \$60 per week. Within these limits he may make his own choices. Thus it may easily chance that a systematic bias will so affect the choice of subjects that the persons actually questioned will not be truly representative of the class to which they presumably conform. A method known as *area sampling* which removes this possibility will be described in Chapter 25. Whatever the explanation may prove to have been in this particular case, the results of the 1948 election call attention to the need for frequent re-examination of the premises upon which a system of stratified sampling is built, especially when social factors that are likely to change with the passage of time are involved.

DEFINING THE UNIVERSE TO BE SAMPLED

In the field of mental testing in general and particularly in the case of intelligence testing there has been an unfortunate tendency to assign to the tests a degree of generality which they do not, in fact, possess. As was pointed out at the beginning of this chapter, it is not easy to determine whether or not bias exists in the selection of items for a test designed to be a sample of some behavioral universe designated as "intelligence," "introversion-extroversion," or the like, since in the very nature of the case such universes are infinite in size and indeterminate as to boundaries. But this is not the case with respect to the universe of subjects for whom the test is to be deemed suitable. This universe should always be regarded as of finite size and with fairly definite boundaries. True, it may later be found that the boundaries may be safely extended so as to include a larger number of cases within the range for whom the test and the normative standards are appropriate. But such an extension should be based upon empirical evidence, not upon unproven assumptions.

In selecting his subjects for the establishment of normative standards, the experimenter will do well to fit his pattern to his cloth. If time and funds permit, he may undertake a more ambitious program in which, for example, an entire state or group of states is taken as the universe from which his sample is to be drawn. If his resources are smaller, it is far better for him to limit the range for which his test is known to be

appropriate and center his efforts upon securing an unbiased sample of a more restricted universe. Of course the geographical factor is not the only basis upon which a universe may be restricted. Age, education, occupation of subjects or, in the case of children, of their fathers, racial or nationality background, income level, or other factors may be specified. Whether the universe be small or large is not of great consequence provided that its characteristics are clearly defined and the sampling of subjects used in reporting standards is selected in such a way as to ensure against bias. As a matter of fact, a careful study of the performance of a well-selected sample of a small and relatively homogeneous universe is often more informative than are the standards derived from one of such diverse composition that meaning is lost for lack of specificity. For example, height standards derived from all the people in the United States regardless of age or sex would have little value. It may well be that an intelligence test planned for use only with children from Taussig's group V (see footnote on page 116), with normative standards for such a group very carefully determined with respect both to mean and to variability would have greater practical and scientific usefulness than many of those now available in which precision of meaning has been sacrificed to secure a wider area of reference—a shotgun instead of a rifle.

As Cantril (1944) has so admirably shown in his studies of the modern public opinion poll, it is none too easy to secure an unbiased sample even for a finite population with clearly defined limits about the characteristics of which something is known. There is no doubt whatever that the marked differences so often found in the apparent standing of the same individual on two or more tests designed for the same purpose are as likely to be the result of differential bias in the reference samples with which his performance is compared as to differences in the manifest content of the tests or to variations in his interest and effort on the two occasions. It is not essential that any one particular universe should always be chosen for sampling, but it *is* essential that the character and scope of whatever universe is selected should be described as exactly as possible, and that every effort be made to ensure that the sample chosen shall be as representative of that universe as it is possible to make it.

THE QUESTION OF SAMPLING IN TEST INTERPRETATION

It may seem at first that what has been said so far in this chapter is intended only for the test maker and has no significance for the test user, who deals only with the finished product—the completed test or scale and the list of reference standards with which he compares his

results. Nothing could be further from the fact. Tests, even the best of them, are by no means foolproof. A telephone can be used as effectively by a person who has no understanding of the principles underlying its construction as by a highly skilled electrical engineer. Unfortunately there are many people who erroneously assume that testing has been brought to a point where the test user is equally freed from the necessity for understanding the instruments he employs. All he need do, they think, is to follow the instructions faithfully and accurately and accept the obtained results at their face value. The mental examiner is thus regarded as a technician rather than as a clinician or a scientist.

The reason for this point of view is largely to be found in the history of the testing movement. In Chapter 5 we saw how the widespread need for the classification of individuals according to ability led to the use of instruments that were by no means perfected, and to a blind confidence in the results obtained by their use that was as much the result of wishful thinking as of objectively determined facts. The need for testing was so great and the apparent nature of the testing devices was so deceptively simple that many people with little or no training in scientific method began to use them freely. Under the conditions then existent, this was not entirely undesirable, for it gave an initial impetus to the testing movement without which it might easily have been stillborn. Moreover, not even the experts of that period were fully aware either of the implicit assumptions upon which the tests which they developed were based or of the limitations and hazards arising out of these assumptions. The use of statistical techniques for the refinement of testing procedures and in the evaluation of test results was not well understood nor was its importance appreciated.

That time is now safely behind us. The number of testing devices on the market has multiplied almost beyond belief, and the number of persons who are devoting all or part of their time to testing has increased proportionately. Inevitably it has occurred that many of the tests are incapable of yielding the results which have been claimed for them and that many of the testers are poorly equipped for the work they are called upon to do. Nevertheless, so much progress has been made toward perfecting the tests that a proper selection from among those available and an intelligent application of their results in the understanding and guidance of human behavior by one who understands the principles involved can be of far greater value than even the early enthusiasts dared to hope. But such understanding is essential. Tests are on their way to becoming instruments of precision, but they have by no means reached the "telephone stage" where they can safely be used even by the ignorant.

Not only those who administer tests which they do not themselves devise but those who make use of the results of tests which they do not

themselves administer should have at least enough understanding of the basic assumptions underlying test construction to enable them to make reasonable judgments as to the significance of these results in individual cases. The principles of sampling, it should be remembered, apply to the establishment of opinions and beliefs as well as to the construction of tests. Individual experience in any field, let us say with the use of tests for certain practical purposes, is merely a single sample of what such tests may contribute for such purposes. As a sample it is subject to bias. Indeed, because of the manner of its selection, it is particularly liable to bias.¹³ The person who has chanced to find—as all will sometimes find—that the results of a particular test given to an individual person at a particular time have proved to be misleading, is too likely to regard this single instance as a representative sample of the universe from which it is drawn. He may even extend the limits of that universe to include all tests, not just the one involved. He may ignore factors which set limits to the universe, such as the competence of the person who gave the test, or the physical or emotional condition of the subject at the time of testing. He may overlook the fact that the sample of subjects used in obtaining the reference standards is always a sample of a finite universe to which the subject may not properly belong.¹⁴ An appreciation of the principles of sampling and a recognition of the likelihood of errors in sampling in individual instances should do much to correct bias in judgment, not only in the area of tests and measurements but in other situations as well.

What the practical worker needs to know before he can advantageously proceed to the interpretation and application of the results of tests is *the character of the universe of which these results may be regarded as samples*. As these pages have repeatedly stressed, this universe must be defined both in terms of the total area of behavior of which the test is designed to be a sample and in terms of the universe of individuals whose behavior is sampled. Such a definition calls for insight as well as for specific information. Both can be gained, but neither is likely to be had without effort.

¹³ No two individuals have had completely identical experiences. Likewise, no two individuals are likely to take identical attitudes with respect to experiences which to an outsider appear closely similar. Thus with the passage of time, attitudes tend to become increasingly individualized (biased) because each new event is viewed in the light of all that have preceded it.

¹⁴ For example, a child whose manual dexterity has been affected by cerebral birth palsy should not be regarded as a sample of a universe of normal children. It would be wholly unsafe to judge his level of intellectual development on the basis of a "performance" test scored in terms of speed, or a group test in which writing or marking is required under rigidly enforced time limits.

The Analysis and Selection of Test Items

TRIDIMENSIONAL ASPECT OF

MOST BEHAVIOR TRAITS

Ordinary observation indicates that the forms of behavior subsumed under any generalized trait name vary in a quantitatively definable manner along more than one coordinate. Although a more elaborate analysis might well indicate a greater number of variable factors, we may note three as outstanding.¹ These are (1) variations in the *kind of conditions* most likely to elicit a display of the behavior in question, (2) variations in the *number of different conditions* under which it is likely to be shown, and (3) variations in the *degree or intensity* of its manifestations as indicated by the difficulty of the tasks with which the individual is able to cope.²

Suppose, for example, we are interested in appraising the ability to influence the behavior of others, generally known as "leadership." We note at once that individuals differ both with respect to the kind of

¹ This analysis has some resemblance to but is not identical with Thorndike's (1926) three "dimensions" of intelligence, viz.; altitude, range, and area.

² Differences in the manner in which the trait is displayed also exist, but I have not included these in the list of basic variables because they are of importance only to the extent that they may render the behavior more difficult to classify. If two individuals are able to deal with the same kind and number of situations *with equal effectiveness*, it is of little consequence whether or not they make use of the same methods for doing so. But it is unquestionably true that an observer who is completely unfamiliar with the method employed by one of the two may fail to follow his line of reasoning and erroneously suppose that his success was merely the result of lucky accident.

However, in comparing the effectiveness of two methods it is necessary to avoid too superficial a view. Dictatorship, observed at short range, may seem a highly effective method of control, but history tells a different story. For the dictator, however successful he may be for a time, almost invariably arouses antagonism which in the end results in his downfall and the disintegration of the organization he built up. The child who uses his fingers as a calculating machine may be marked "perfect" on all his arithmetic papers in the primary grades but unless he changes his methods he is not likely to maintain his standing indefinitely.

people whom they influence most readily and, as a corollary, with respect to the kind of situations in which their leadership is usually manifested. Here is a man who takes the lead in the various activities of his church. He is likely to be the chairman of church committees, he is active in raising funds for church support, and his views on matters of church policy always receive respectful attention and in most cases are accepted. But in affairs outside the church he plays only a minor role. He belongs to two or three clubs but has never held office in any of them. He once ran for membership on the board of education but received only a few votes.

Much the same state of affairs exists in the case of a second man, except that his leadership is confined almost entirely to the members of a ring of gangsters whom he dominates in part by force but chiefly by superior daring and cleverness. The two men resemble each other in the narrow range of circumstances under which their leadership is displayed in spite of the marked contrast in the nature of these circumstances. And if we are interested purely in the question of leadership as such, we must not apply to our data such extraneous criteria as the social desirability of the behavior in question. Within another frame of reference this may be a highly important question, but in this case it can only confuse the issue. Nevertheless, the fact that a given trait may be manifested in any of a wide variety of different circumstances is a matter of much importance for the test constructor, as well as for the practical user of tests. For if the tasks which are set fall within too narrow a range, if they are too much alike in kind, there is grave danger that those persons whose ability is most readily manifested in other aspects or areas of the field in question may be seriously misjudged. Thus, in the example just given, if we were to confuse leadership with social desirability and choose our leadership items only from areas that meet with social acceptance, the ability of the gangster would be greatly underrated. In like manner, an intelligence test which does not provide for the exercise of the individual's mental powers in a reasonably wide variety of situations will give too favorable results for those whose strong points are given an opportunity for display but whose weaknesses are not exposed, while the reverse will be true in other cases.

The variety or range of situations in which a trait is manifested is also important. Here is a third man who differs from both of those previously described. Almost regardless of the conditions in which he chances to be placed, he quickly becomes the dominant personality. It is he who organizes a bridge tournament when his train is stalled by a flood. He is active in church work, president of the parent-teacher association, secretary of his golf club. He has held a number of public offices in the small town in which he resides. When they are in difficulty, his

friends are likely to come to him for advice. While his sphere of influence is local rather than national in its extent, it nevertheless covers a fairly wide range of persons and circumstances. Certainly its scope is much broader than that of either of the two persons first mentioned. Yet the situations in which his leadership is displayed are for the most part neither very elaborate nor highly difficult to handle. The groups he leads are for the most part small; the decisions he has to make do not, as a rule, involve complicated problems or carefully planned campaigns. In this respect he differs from the leader of industry, the social reformer, or the statesman. The leadership of these persons may be exercised in many fields or be confined to the single area in which they have won renown, but in any case they have demonstrated their ability to deal with problems which most men would find it impossible to handle.

The same principle applies in such a field as intelligence. Some people are able to act intelligently in many situations; others in only a few. Some excel, that is, they show their greatest ability, in one field; others in another; still others in a third. Whether the scope of their intelligence is wide or narrow, there will also be differences in its degree or level as indicated by the difficulty of the intellectual tasks which they are able to master.

All this has a definite bearing upon the selection of items to be included in a test of any mental trait. If bias is to be avoided, care must be taken to sample all the major areas in which the trait in question is likely to be displayed, and to include within each of the areas a series of items which vary in difficulty from those which can be passed by persons who manifest the trait to only a very slight degree, to those with which only a small number of the highly gifted will be able to cope.

These principles also have an important bearing upon the selection of tests to be used in the appraisal of individuals and on the interpretation of test results. If biased conclusions are to be avoided, it is necessary to make sure that the content of the tests used is appropriate to the experiential background of the individual who is tested. This does not mean that only those tests should be given him which stress the areas in which he is known to excel. Such a procedure would be as misleading as is that of the grocer who carefully sorts out the best strawberries to put on top of the box. It does mean, however, that no person shall be judged on the basis of tasks that he has had no opportunity to master or for which his opportunities have been distinctly more limited than those of the group with which he is to be compared.³

³ Such comparisons are, of course, entirely legitimate when the purpose of the experiment is not to compare the abilities of different persons but to appraise the effect of their differing experience.

THE SELECTION OF TEST ITEMS

It will be evident at once that unaided human judgment is not a safe guide for meeting either of the two criteria just mentioned. At the outset, it is true, the selection of items to be tried out will be determined by the experimenter's acquaintance with the results obtained by his predecessors and by the observations and hypotheses with which he begins his work. The more clearly the problem has been defined in advance and the more complete the preliminary analysis of the behavior which it is desired to predict with reference to the areas in which it is likely to be manifested, the easier, as a rule, will be the subsequent task of drawing up a preliminary series of sample items for trial and validation. If the characteristic to be tested is the degree of ability, aptitude, or skill possessed by a subject in any one of a wide variety of specified areas, and the worker has a reasonably clear idea of the range and variety of situations in which the ability in question is likely to be manifested, it will usually be possible to devise a series of tasks which are judged to form an approximately representative sample of the total by including a suitable proportion from each of its subareas, so chosen as to cover the entire range of difficulty levels appropriate to the subjects for whom the test is devised. In the early days of testing, these items would then have been arranged according to some preconceived system of order, such as kind of task or apparent difficulty, and the result would then have been called a "test" of mechanical ability or whatever skill the test constructor had in mind. But modern research has shown that such armchair procedures usually lead to methods that not only are wasteful of time but may be downright misleading. The tasks drawn up in this way constitute a necessary starting point, a trial series, the usefulness of which is still to be determined. Careful statistical evaluation of each of the separate items will usually show that some are unnecessary since, although they may appear to be different, they actually merely duplicate the information given by others. Some may have been so poorly formulated that different subjects are likely to interpret them in varying ways.⁴ Still others may show so little relationship to the remainder of the series, or to such other evidence as can be obtained with respect to the standing of the subjects

⁴ Some attempts have been made to utilize the fact that certain types of instruction or certain kinds of stimuli will mean different things for different people, according to their past experiences, interests, and attitudes. Because of these differences in meaning, such persons will react in varying ways to the same set of instructions, and if these differences in type of response are sufficiently clear cut so that there is usually no difficulty in deciding which of the two or more possible meanings the subject has selected as the true one, the use of such bivalent instructions may yield information of a kind that could not readily be secured by more conventional methods

on the trait in question, as to raise serious doubts concerning the wisdom of retaining them. Some, however, will in all probability prove to be sufficiently reliable and valid to be worth retaining without material change, and these will serve at least as a nucleus to which other items may later be added if it is found necessary.

The first step in the evaluation of sample tasks or items is to make sure that the wording is free from unintentional ambiguities. This is by no means a simple task. As Cantril and his associates (1944) have shown, the formulation of suitable questions is a task beset by many difficulties which even those with much experience in test construction too frequently overlook. While the hazards arising from poorly worded questions are undoubtedly greatest when the method of direct questioning with respect to beliefs, attitudes, personal experiences, and the like is depended upon, as is usual in the so-called "personality inventories," "attitude tests," "interest tests," and the like, they are by no means rare in the forms of instruction used in tests of ability or in the questions asked when the aim is to test general information or knowledge of some special field. Even when a question seems to its originators perfectly clear cut and straightforward and the responses of the subjects do not suggest any misunderstanding or confusion until further analysis is made, such misunderstandings may nevertheless exist. Cantril reports the responses to this question which was employed by the Office of Public Opinion Research during World War II as a means of gauging public opinion toward a negotiated peace: "If the German army overthrew Hitler and then offered to stop the war and discuss peace terms with the Allies, would you favor or oppose accepting the offer of the German army?" Although the wording of the question seems clear enough and there was nothing in the distribution of affirmative and negative replies to suggest that anything was wrong, when the members of a selected sample of respondents were asked to state in their own words just what the question meant to them, it was found that the majority of those who favored acceptance had identified the German army with the German people, whereas those who opposed accepting such an offer had understood the question to refer only to the military powers in whom they had as little confidence as in Hitler himself. This is a neat example of the subtle manner in which unsuspected semantic differences may give a totally different significance to replies from that which would normally be

(Goodenough, 1942, 1946). But in these cases the instructions should be regarded as ambivalent rather than ambiguous, and the responses appropriate to each of the possible meanings should be analyzed as carefully for objectivity and significance as is done when only one meaning is intended.

assumed on the basis of straightforward interpretation of them. In Cantril's chapter entitled "Meaning of Questions" he lists and discusses eleven different types of common errors in the formulation of questions, most of which apply as much to questions used in mental tests as to those of the public opinion poll from which his examples are derived. This chapter is one of the most searching dissertations on questionnaire construction that has appeared in the literature. It should be required reading for everyone who constructs or uses tests of the questionnaire type.

It is thus unsafe to depend upon personal judgment in assuming that a given question is free from ambiguities or other types of error that alter the significance of the replies to it. Cantril's method of testing the form of questions—by asking a selected sample of the kind of people with whom the test or questionnaire is to be used to explain just what they understand by it—is useful when the subjects are of an age and intellectual level that will make such explanations possible. Young children or persons with very inadequate command of English may not be able to do this, although many who have not tried it will be surprised to find how far down the scale of intelligence such explanations become possible. When direct examination is not feasible, the proposed questions (or test instructions) should in any case be gone over critically for clearness and precision of meaning by a number of persons who have had experience in the field.

When a preliminary series of sample tasks has been formulated, the next step is to eliminate those which prove to be either unnecessary or useless for the purpose at hand. In order to do this a criterion is required. And inasmuch as the significance of the test will depend largely upon the nature and adequacy of the criterion used for standardizing and validating it, the choice of a criterion merits far more attention than has often been given it. The extenuation offered for the use of poor criterion measures has commonly been that no better data are available and that the reason for constructing a new test is the absence of any valid and objective device for appraising the trait in question. This, however, is not always true. Sometimes the available methods are too cumbersome or time consuming for general use, and a shorter or more convenient method is needed. In such cases, if the long method is known to be satisfactory, enough data may be secured in this way to serve as a criterion for standardizing the shorter method, even though it is not feasible to use the former for more extensive work. Sometimes a criterion becomes available only after the need for a test has passed. This is often the case when a test of aptitude for a particular kind of work is wanted. Success on a job is a practical indication of aptitude for that work, but one that

cannot be had until it is too late for use in the selection of workers. The degree of success of experienced workers, however, may be a valuable criterion to be used in the derivation of a test to be given to prospective candidates for the kind of position in question. If at all possible, the test should be given to the members of the criterion group *at the time they apply for the position*, since scores earned after a period of experience on the job may not correspond to those which would have been obtained at the outset. Moreover, if only those who have been able to retain their positions for a period of time are tested, the range of ability available for comparison will be greatly restricted. However, if conditions make it unfeasible to test first and wait until the degree of success has been demonstrated by actual trial, the amount of experience of each worker in the criterion group should be kept the same.⁵

If an outside criterion can be had which is *free from bias*, that is, some quantitatively expressed sign which represents the characteristic to be appraised in a fair and representative manner, chance fluctuations in the magnitude of the scores obtained by its use are of less consequence than might be supposed. Low reliability of a criterion (in the statistical sense of the word) which is otherwise valid and unbiased means only that more data will be required to cancel out the effect of chance errors. But a biased criterion, one that does not yield a truly representative picture of the universe for which it stands, will inevitably lead to an unbalanced test which will overrate some individuals and underrate others. If, for example, the criterion used in the selection of items to be included in a test designed to appraise emotional maturity in elementary school children should be the judgment of a single teacher as to the level of emotional maturity attained by her pupils, and it chanced that this teacher regarded obedience to classroom regulations as the best sign of the trait in question, it is obvious that a series of test items chosen to conform to this criterion could not yield an unbiased measure of the universe which it was desired to appraise. But if it were possible to identify a sign or a series of signs which showed greater than a chance tendency to point in the desired direction, and no consistent tendency to point in any other, even though they might be found to be very undependable indicators of the trait because of their lack of stability, they would still constitute a better criterion for the development of a new test than would a more stable but biased measure. By securing a suffi-

⁵ Partial correlation is sometimes used as a means of holding experience constant when experimental control does not appear feasible. Unless the relation between test score and length of time on the job is so slight that errors from this source are of little consequence, the method is of questionable validity since a time curve is rarely a straight-line function. If this method is resorted to, the correlation ratio should always be used unless rectilinear regression can be established. (See Chapter 17.)

ciently large amount of data it is possible to develop a scale which will yield results that are far more stable ("reliable") than is the criterion used for the selection of items, but no amount of data will compensate for bias in the criterion. Thus it behooves both the test maker and the test user to give careful attention to the validity of the criteria used in the standardization of the tests with which they deal. It will not infrequently be found that the test maker has begun by *assuming* the validity of his criterion and has later reported on the "validity" of his test in terms of its agreement with the original criterion. If his initial assumption was sound, this procedure is of course justified; but if the criterion was biased, the procedure merely demonstrates that the test is biased in a similar manner.

It not infrequently happens, however, that no suitable criterion can be found which can be depended upon to be free from bias. Under these circumstances test makers frequently depend upon the internal consistency of the items as an indication of their value for the scale as a whole. Various procedures, such as the biserial correlation of each item with the sum of all the others, the proportionate number of responses of a given type made by the upper and lower quartiles of the distribution of subjects, as well as a number of graphic methods, have been used both for the elimination of items which are judged to be of little value and for the assigning of weights to those that are retained (Guilford, 1936). A special device for determining whether or not a given item should be regarded as part of a universe (either as a sample or as a sign), which is of particular value when the items in their original form are qualitative rather than quantitative in nature, such as statements of opinion, descriptions of behavior or appearance, and so on, has been outlined by Guttman (1944). The argument is relatively simple and straightforward. It says, in effect, that items are scalable, i.e., susceptible of reduction to numerical values, if they can be arranged in such a way that from the form of response given to certain ones, the responses to others can be inferred with practical certainty.⁶ If the universe to be appraised is simple, that is, varying only with respect to degree, Guttman's method is useful; but if variations in range or scope are also to be taken into account, some modification of the procedure would appear to be needed. Guttman notes that data may be scalable (according to his method of determination) for one population at a given time but not for another population or for the same population at another time, or the scale

⁶ Because of chance errors, such an inference can rarely be made with absolute certainty. Guttman states that in practice an agreement of 85 per cent or more may be considered perfect in the sense that only chance factors are likely to have been operative.

values may change. He notes also that in practically any set of data, individual cases will be found whose responses do not accord with the principle on which the scale values were determined. The inferences which are valid for the majority do not hold good for these persons. Guttman suggests that from a study of such cases, valuable information with respect to the factors influencing the choice of responses may be gained.

In summary, then, we note that two very different concepts have been operative with respect to the selection of items to be included in a test battery designed for the appraisal of individuals with respect to some specified form of ability or behavioral tendency.⁷ The first assumes that the universe to be appraised must be predefined in terms of some reasonably objective criterion other than the group of items to be included in the test itself. The value of an item is then determined on the basis of its contribution to the prediction of the criterion. The use of an external criterion makes it possible to define a universe in sufficiently broad terms to allow for a good deal of individual variation in the manner in which a given characteristic or ability is manifested and to take account of the range or scope of conditions under which it is displayed as well as of the degree or level of accomplishment within a given area. Its greatest hazard lies in the possibility that the criterion may be biased. A test which is developed on the basis of a biased criterion cannot yield a fair appraisal of the individuals with whom it is to be used unless the test constructor has been cognizant of the extent and direction of the bias and has taken steps to allow for it when standardizing his test.

The difficulty of finding suitable criteria has led many test makers to the view that the organization and internal relationships of items that presumably competent judges select as probable samples (or signs) of the characteristic in question are likely to afford the best available evidence of the suitability of a given item for the purpose at hand. An item which is found to have something in common with the others, as determined either by some form of correlational procedure or by its conformity to an organizational hierarchy as proposed by Guttman, is judged to be useful; others are eliminated. Weights may be assigned to the retained items on the basis of their correlation with each other or with the total or on the basis of their place in the hierarchy.⁸

⁷ An "ability," as we have used the term, refers to the level of attainment which the individual is capable of reaching under reasonably favorable conditions and when highly motivated; a "behavioral tendency" refers to his usual or most typical conduct.

⁸ Place in the hierarchy is based upon the number of other items to which the responses may be predicted from that given to the stated item. For example, it may be assumed that if a respondent states that he will attend college no matter how good a job may be offered him he will also state that he will attend even if a fairly good job

Accordingly, when this method is employed for the selection and weighting of items, any which are unrelated to the others will be eliminated, even though, if the facts were known, their contribution to the universe which it is desired to appraise is of special importance just because it is not duplicated by any of the rest. Thus the use of internal consistency when employed for the selection of test items is likely to result in a test of narrow rather than broad significance, and if it so chances that the person who made the original choice of items to be tried out was biased⁹ in his view of the nature and organization of the trait which he desired to appraise, his own bias will be reflected in the correlations of the separate items with the total, since certain areas will be represented by more than the appropriate number of items, and success or failure on one of these items will accordingly show a closer agreement with the whole than will the performance on another of a type which has a smaller representation in the total. All this makes for a tendency toward the selection of items which duplicate each other wholly or in part. As a result of this duplication, chance errors will tend to cancel each other and thus increase stability of measurement and produce greater specificity of meaning. Specificity has its advantages, but if the test maker has begun his work with the hope of securing a measure of some rather broadly defined universe and fails to realize that his method of standardization has been such as to restrict the meaning of his test to comparatively narrow limits, the result may be misleading.

THE ANALYSIS AND SELECTION OF ITEMS IN TESTS OF THE LIMITED-RESPONSE TYPE

In group testing, the subjects are not, as a rule, permitted to respond to questions in their own words because of the time required for reading the papers and the likelihood that different examiners will employ different standards in grading the replies. The usual practice is to provide a series of alternative answers from which the subject selects the one that he judges to be the most nearly correct. The preparation of an examination of this kind is not an easy task. Not only the wording of the ques-

is offered, and also that he will attend if he cannot get any kind of a job. However, the weight given is determined on the basis of the organization of the responses actually found for a given group of subjects, not upon any assumption as to how they should be organized. If a fairly large number of items is included in the series, it will ordinarily be found that some responses can be inferred from others with a satisfactory degree of certainty, even though they deal with matters that superficially appear quite different.

⁹ That is, if he held a distinctly different opinion from that of most other competent persons. Trait names, like other linguistic symbols, are determined by popular usage.

tions to be answered but the selection and wording of the responses from which the subject is to make his choice will have an important bearing upon the results. A question that might otherwise be very useful can lose all or most of its discriminative value (a) if some or all of the list of responses are ambiguous in meaning, (b) if all or most of the incorrect responses are so manifestly wrong that the majority of the subjects will be able to choose the right one by a simple process of elimination, (c) if previous questions have been so worded as to provide cues that will aid in the answering of subsequent ones, (d) if there are internal cues in the question itself that suggest the answer to be chosen (such as the use of "a" or "an" or of a singular or plural verb with which not all the indicated answers agree). These and other similar points should be considered with much care when drawing up the preliminary form of a test for later trial and standardization. Equal care should be exercised in preparing an objective examination for ordinary classroom use. If at all possible, the questions should be gone over in advance by one or more persons who have had considerable experience in test construction. In addition, several students or other persons who are reasonably representative of the group of subjects for whom the test is designed should be asked to take the test. They should be requested to give their reasons for each choice of response and to report any difficulties they may experience in understanding the questions or in deciding upon the response to be chosen.

The form and arrangement of the items used in objective examinations differ considerably. The "multiple-choice" arrangement, which is widely used, consists of a question or statement followed by a list of responses among which the subject indicates his choice in some designated way. Examples are:

Spare is a term used in: tennis; bowling; football; cribbage

If you were asked your opinion of someone whom you do not know, which of the following answers would you give?

- I think he is all right.
- I will try to get acquainted with him.
- I don't know him and so cannot tell you.

The number of stated alternatives usually varies from two to five or, rarely, more than five. Four is a common number. Both the so-called "matching test" and the "true-false" test can best be considered as variant forms of the multiple-choice method. In the matching test the subject is presented with two columns of items. Each of the items in one column is related to one of the items in the opposite column in some specified way. For example, the left-hand column may be made up of names of

authors; that on the right may contain names of their books, arranged, of course, in random order. Each author's name will be preceded by a number. The subject is asked to match the two lists by inserting the author's number in a blank space preceding the name of his book.¹⁰ The true-false examination consists merely of a series of statements which the subject is required to mark as true or false. It is thus merely a multiple-choice type of questionnaire with only two alternatives for each question.

The "completion test" differs slightly from the other types in form, though its statistical treatment is the same. It consists of a number of sentences or short paragraphs from which one or more words have been omitted, and which the subject is required to supply from the context. As a rule, in tests of this kind only the officially correct answers need be subjected to statistical treatment. But if, as sometimes happens, two or more different completions occur in sufficient numbers to warrant it, they may be handled in the same way as multiple-choice responses.¹¹

The steps involved in the selection and evaluation of a series of items that are to comprise a test of some predesignated ability may be outlined briefly as follows:¹²

1. Prepare a clear and unambiguous definition of the universe of which the test items are designed to be samples.¹³ This definition should include an analysis of the universe into its major components. For example, it might be desired to construct an examination to measure the extent of knowledge of English history possessed by college students. It is necessary to decide first whether a simple knowledge of facts or an understanding of the relationship of these facts to one another is to be stressed. The field to be covered should then be subdivided into its major units or topics and the proportionate weight to be given to each should be determined.

2. Prepare a series of tasks or questions corresponding to this analysis. The form in which these questions are couched will usually depend on their character. Some questions lend themselves best to the matching

¹⁰ One of the columns should always contain several more items than the other in order that the matching of the last items may not be automatically determined by elimination of all the others.

¹¹ Paterson (1925) and more recently Ross (1941) have given excellent suggestions for the preparation of examinations of this kind. Adkins (1947), whose experience in the preparation of civil service examinations has particularly qualified her for the task, has described a number of ways of evaluating test items and of developing criteria by which their worth for a designated purpose may be judged.

¹² It should be noted that this outline is concerned only with the selection of test items. It does not cover the later problems of scale construction or the establishment of normative standards.

¹³ Since the character of the universe is predesignated rather than emergent, it follows that the method of sampling rather than the method of signs will be utilized.

technique; others, to the multiple-choice form; and still others, to completion exercises. The number of items in the original list should be greater than the number that the examiner hopes to retain since some will almost certainly be rejected when the item analysis is made.

3. After the manner described in a preceding paragraph, have the questions examined for clearness and freedom from unintentional ambiguities by at least one "expert" and several naïve subjects whose level of competence is similar to that of the persons for whom the test is designed. Make any revisions that seem called for.

4. If the test is planned for extensive use with large groups, a good deal of time will eventually be saved by basing the item analysis on the results obtained for a preliminary group of subjects selected to be representative of the population for whom the test is intended and sufficient in number to lend stability to the data. The test can then be put directly into its final form. If the test is intended only for temporary or local use, the item analysis may be deferred until after the test has been given. The papers will then need to be rescored on the basis of the corrections that have been made.

5. If possible, select a criterion outside the test itself for determining the validity of the items. Examine this criterion carefully with reference to its comprehensiveness and freedom from bias. Sometimes it is possible to make use of objective tests for this purpose. Often it will be necessary to depend upon the test maker's acquaintance with the proposed criterion and with the factors likely to influence individual standing with respect to it. All the subjects whose test performance is to be used in carrying out the item analysis must also be tested by means of the selected criterion.

6. Each type of response to each item,¹⁴ including the presumably incorrect as well as those which the examiner has assumed to be correct, should then be compared separately with the score on the criterion. It sometimes happens that the choice of some particular incorrect response will be found to have greater diagnostic significance than the selection of the correct response. Particularly in the case of personality inventories, rating scales, and the like, armchair judgment as to the "best" response may be very undependable. The most usual way of making the comparison is by means of biserial r . (See Chapter 17.) An alternative method that is also widely used consists in arranging the subjects in serial order according to their scores on the criterion and using the difference between the percentage of those in the top 25 per cent who select a given response and the corresponding percentage of those in the bottom

¹⁴ In the case of purely factual questions, separate analysis of the definitely incorrect responses may be omitted.

25 per cent who choose the same response as an indicator of its discriminative value. Regardless of the method used, a critical point is chosen to serve as a basis for retaining or discarding an item. If the biserial r method has been used, this point will be expressed in terms of the likelihood that the obtained r truly exceeds zero; that is, in terms of the probability that in other samples drawn from the same population the r 's would have the same sign (plus or minus) as that found for this sample. If the second method has been used, some point in the distribution of the t statistic (see Chapter 16) will be chosen as the dividing line.¹⁵ All items in which the result obtained for at least one of the set responses equals or exceeds this value will be retained as a part of the final test; the remainder will be discarded. The level of significance required for the retention of a single test item is much lower than is commonly thought necessary for the refutation of the null hypothesis (see Chapter 16) where the results of an entire experiment are concerned. Most people have found that a value of r or of t that reaches the 25 per cent level of confidence is sufficient to justify the inclusion of an item in a test made up of not fewer than 50 equally weighted items. A higher degree of dependability of the individual items is needed when the total number included in the test is small.

Negative as well as positive scores may be given. If it chanced that the choice of the officially "right" answer does not discriminate between strong and weak subjects but that the selection of a particular one of the incorrect responses does prove to be discriminative (and in the proper direction), scoring may be done on the basis of the incorrect response and the item given a negative weight.¹⁶ The score on the entire test then

¹⁵ The term "level of significance" is commonly used to denote the number of chances in 100 that the direction or general tenor of a statistical finding will *not* be duplicated if the experiment is repeated under like conditions but with a different sampling of subjects. In the case under consideration, if r or t is found to reach the 1 per cent level of significance (or "confidence"), it would mean that there is not over 1 chance in 100 that in further trials r would not have the same sign or that the difference between the proportion of the upper and lower quartiles of the subjects who give the response in question would not lie in the same direction as was found for the present sample.

¹⁶ In most cases only one of the set responses will be used as a basis for scoring the item; others are disregarded. The reason for evaluating all of the responses is twofold: first, because, as indicated above, one of the incorrect responses may prove to be a better (more discriminative) basis for scoring than the right response is found to be, and, second, because such an analysis provides a further guide to lack of clarity in the instructions or to ambiguity in the wording of the questions. Theoretically, only the one "correct" response should receive a positive weight when the item analysis is made; all others should be negatively weighted. But it may sometimes happen that the more able subjects select one of the incorrect responses almost or quite as often as they choose the right one, with the result that both of these are found to have a sufficiently high discriminative value to warrant their inclusion with a positive

becomes the difference between the sum of the scores on the items positively weighted and that on those given negative weights.

7. When no outside criterion is available, the test is first scored on the basis of the total number of responses assumed to be correct. This score is then used as a criterion with which the individual responses to the separate items are compared in the same manner as is used when an outside criterion can be had.

8. The question as to whether or not all the items that are retained should be given equal weight in determining the final score, or whether they should be weighted differentially according to their discriminative value and freedom from overlapping,¹⁷ has never been answered satisfactorily. Theoretically, a system of differentiated weighting should be preferable. In practice, the advantages of such weighting have rarely been found to be great enough to justify the labor involved.¹⁸ When the total number of items is small, differential weighting may sometimes prove advantageous, but with the number of items usually included in tests of this kind, the correlation between the scores obtained when all the items are given an equal weight of 1 and that found when differential weights have been assigned to them has usually been well over +.95 and sometimes as high as +.99. Inasmuch as not only the time required for the original evaluation but the labor involved in scoring the tests later on will be considerably increased by the use of differential weighting, we may well question the practical justification for its use in the greater number of instances.

9. Finally it should be emphasized that the use of the total score as a criterion for evaluating the separate items can best be justified when the test is of a kind that deals chiefly with the knowledge of facts or the possession of specified skills. When the nature of the universe is not

weighting. Had only the "right" responses been evaluated, this difficulty might never have been noted. In most cases it is wise to discard items of this kind on the basis of their apparent ambiguity or the possibly controversial nature of the issue involved.

It should be unnecessary to point out that what has just been said does not apply to measures of preference or choice which do not involve right or wrong answers and where the responses are classified as to type rather than graded as to quality.

¹⁷ If two or more test items are related in such a way that success or failure on one may be predicted from success or failure on the other with a sufficiently high degree of assurance, the inclusion of both in the same test series is wasteful of time. However, the amount of time and labor required to determine all the intercorrelations of the items in a multiple-itemed test is so great that such calculations are rarely made unless a factorial analysis is to be carried out.

¹⁸ If the original division of the universe represented by the test into its major components or subtopics has been well thought out, and the number of items under each of these divisions has been made roughly proportional to their importance in the total complex, this in itself amounts to a kind of weighting, although it is likely to be based upon subjective rather than objective evidence.

clearly defined, bias in the choice of items is likely to occur; and this bias will not merely be reflected but be accentuated when the test score becomes the criterion. Its use, therefore, is never wholly satisfactory, but it may nevertheless be preferred if the only available external criteria are even less satisfactory. If the hazards and advantages of each of the two methods are clearly understood, an intelligent choice between them can more easily be made.

SOME PRACTICAL APPLICATIONS

The clinical worker or vocational counselor who attempts to make use of tests in the guidance of human beings should understand something of this relationship between the method employed in standardizing a test and the kind of conclusions that may be drawn from its use. Too often, test users are content with a simple statement of the test's "reliability" and "validity." They fail to realize that such figures may mislead as often as they inform unless other facts needed for interpreting them are known. High self-correlations ("reliability") afford no evidence that the test is either a useful sign or a sufficient sample of the characteristic of which it bears the name. The two may be quite unrelated, or the test may represent only a very limited part or aspect of the trait in question. As we have seen, the latter result is especially likely to occur if the internal consistency of the items was the criterion used for judging their suitability. Tests with narrow meaning are often exceedingly valuable instruments for use in their own limited fields. Their very specificity makes for precise interpretation of the results obtained by their use, provided always that their meaning is known and its limitations are recognized. But this is essential.

High "validity coefficients" (correlations with criteria outside the test itself) may also be misleading. Often such correlations have been obtained from measures which resemble the test so closely that the figures should be looked upon as self-correlations rather than as validating data. Sometimes they are based upon groups of such extremely wide dispersion that even the crudest measure would suffice to arrange the tested individuals in order of merit. A very imperfect measuring rod would indicate that a typical newborn baby, a six-year-old boy, and an adult are not equal in height. It does not take a good reading test to show differences in the reading skill of children whose grade placements run all the way from the second to the eighth grade. Yet many such correlations are reported in the literature and are accepted by those who have not taken the trouble to examine the facts as indications that the tests in question can be depended upon to meet the purposes for which they were designed.

In the seventeenth century there was widespread belief in the principle known as "sympathetic magic." Even today one need not search far to find persons who still cling to some aspects of this belief. The basic principle of sympathetic magic is that some mysterious relationship exists between the symbolic representation of a person or an object and the individual himself. Thus it was believed that making an image of an enemy and then mutilating it in some way would work a corresponding injury upon the person whom the image was supposed to represent, or that a curse might be laid upon a person by mentioning his name with the appropriate incantations.

Modern students of semantics have shown that in spite of present-day emphasis upon scientific objectivity, our word symbols carry with them something of the same framework of magical transference. We call a thing by a certain name and at once it becomes invested for us with certain qualities pertaining to that name. The field of tests and measurements is full of such examples. Someone devises a series of tasks which he calls a "test of leadership." This test is given to a group of high school students. A bashful little girl with horn-rimmed glasses who rarely speaks above a whisper makes the highest score in her class. "Isn't it wonderful?" cries the enthusiastic tester. "No one ever thought of her as a leader!" But an examination of the criteria used in developing the test would have given the explanation, if it were found that the maker of the test believed that the most important characteristic of a leader is his ability to command the respect of others through a wide fund of general information, and if the girl in question chanced to be an omnivorous reader whose very lack of social effectiveness threw her back more exclusively upon reading as a pastime. Calling the device a test of "leadership" could not change its essential nature. The reported correlations with various criteria selected by the author in terms of his own biased judgment might be satisfactorily high, but they would not necessarily represent the kind of relationship likely to be attributed to them by the person who thinks of leadership as a social rather than an intellectual attribute.

Knowledge of the criterion in terms of which a test was developed is thus an important part of the professional equipment of those who wish to use such a test with understanding and profit. It is unsafe to depend solely upon test names or even upon figures purporting to give information about the test's "reliability" and "validity." A careful study of the methods employed for selecting the test items and of the theory which led to the choice of these methods is likely to provide a sounder understanding of the possibilities and limitations of a given test than can possibly be had in any other way.

Units of Measurement

REFERENCE STANDARDS AND INTERPRETATIVE MEASURES

Certain very primitive tribes are said to have no abstract numerical symbols. In the languages of these tribes one word means "two goats," another means "two men," and still another means "two trees," but there is no single word that stands for the abstract quantity, *two*.

In the early days of mental testing much the same situation existed with respect to the expression of test results in a way that could be universally understood without reference to the particular test used. It was possible to say that John Doe had passed sixteen out of twenty-four items included in such and such a series of tests. But whether John Doe was bright or stupid could only be determined by reference to a table of standards for that test. No general interpretative term had as yet been devised by which the meaning of his performance could be made independent of the particular test on which it was earned. For lack of a uniform abstract term which could be applied to all, test scores were accordingly very difficult to handle or to interpret.

In Chapter 4 we saw how Binet and Simon attempted to solve this problem in the case of children by introducing the term *mental age*. Later on, Terman carried the process of abstraction one step further by incorporating into his 1916 revision of Binet's scale the use of the *intelligence quotient*, which had been suggested by Stern and by Kuhlmann some years earlier. By this time there had been proposed a number of other interpretative devices calculated to free the test user from the need of constant reference to a table of standards and to make the results obtained from one test directly comparable with those of another. None of these, however, was destined to attain a popularity equal to that of the two just mentioned. The convenience of these devices for quick and easy interpretation of test results led to their immediate adoption not only by persons who understood their significance and

limitations but also by many others who lacked this knowledge. As a result, many erroneous ideas became current, a good many of which found their way into print and became common beliefs among well-meaning but poorly trained mental examiners and others who made use of their findings. An account of these misconceptions would serve no useful purpose here, but a consideration of the problems involved in passing from a simple item count to more generalized expressions applicable to a wide variety of tests and measurements may provide the basis for a better understanding of these terms and of their limitations and possibilities.

ITEM COUNTS AND "ABSOLUTE SCALING"

In the last chapter it was pointed out that the items included in a test must range in difficulty from very easy to very hard if the various levels of ability attained by different persons¹ are to be satisfactorily appraised. But unless special attention is given to the matter, it is unlikely that the increases in difficulty from one task to the next will be equally spaced. One inch on a scale of length is as long as any other inch. If this were not true, if, for example, the unit of distance lying between the points marked 11 and 12 on a foot rule were only half as long as that between the points marked 4 and 5, the measures of length obtained by the use of so poorly calibrated an instrument would not be easy to interpret. They would not be entirely useless, for the points marked off would still possess the merit of being arranged in rank order in a linear series. It would still be possible to say, "Ten inches on this scale is more than nine inches, seven inches is more than six inches," and so on, but whether or not the differences from one marked point to another were comparable with each other would not be known, nor would it be possible to say with assurance that an object "measuring" ten inches in length on this scale is twice as long as one of five inches.

When a series of test items has been selected, it is not difficult to arrange them in serial order of difficulty by comparing the percentages of subjects who are able to succeed with each. By the use of sufficiently large populations, errors of sampling² can usually be reduced to a

¹ The range of difficulty included in any test or scale will, of course, be adapted to the class of persons for whom the test is intended.

² Two kinds of sampling errors must be reckoned with. First there are the errors arising from the fact that no person is always at his best. Circumstantial factors of various kinds will affect his performance from time to time. The single test is thus a sample of his performances on tasks of this kind. The second has to do with the particular individuals whose performance is tested and who, presumably, constitute a representative sample of a larger population to whom the test is judged to be applicable. If the factors which modify individual performance from time to time are

point at which they are of negligible consequence. The rank order of the items can then be determined with a satisfactory degree of accuracy. But as Thorndike (1926) has shown, the problem does not end here, for tasks may vary in difficulty for any of a number of reasons, not all of which may be pertinent to the characteristic in which the experimenter is interested. Few tasks have been or can be so completely "purified" that they make demands upon only one kind of ability or give occasion for the display of only a single aspect of the personal-social characteristics of the tested individual. Of two items included in a test of intelligence, one may be clearly more difficult than the other as judged by the comparative percentage of persons who are able to succeed with them, but the small number of "passes" in this case may be the result of extraneous factors rather than of intrinsic difficulty. For example, it is known that abstract nouns are harder for children to define than are the names of common objects. This is chiefly a matter of intrinsic difficulty resulting from the basic character of the concepts involved. But among the names of objects there will be some that are quite as unlikely to be known as are many abstract nouns. The difficulty of these words when included in a vocabulary test is largely extrinsic; the objects in question have not chanced to come within the experience of the children. There is, to be sure, an intrinsic factor involved as well, for children of a very inquiring turn of mind will often learn about objects with which they have had no direct experience, through hearing them mentioned by adults, from pictures, and from other indirect sources, while the less able are more dependent upon firsthand contact. In spite of this, it is unquestionably true that the proportion of a given population able to pass a particular item is not an absolute guarantee of its intrinsic difficulty. Evidence for this statement can easily be had. If two groups of moderately contrasted experience are given almost any ordinary test of intelligence and the percentage of each group which passes each item is determined, it will usually be found that the order of difficulty is not identical for both. The differences are usually small if the items have been chosen with the factor of

distributed at random among the group tested, they will tend to cancel each other, but this may not always be true. Hurlock (1924), for example, found that praise or reproof, when used as incentives during the administration of a group intelligence test, had a differential effect upon the performance of different groups, some being more affected by one, others by the other. To the extent that such factors exist, it may not always be safe to assume that the addition of more cases will eliminate sampling errors of the first type.

Errors of the second type have been discussed before. They may be expected to cancel each other in large populations if and when (a) sufficient information is available with respect to the stratification of the characteristic in question within the universe for which the test is designed to make a representative matching possible, and (b) such matching is actually carried out.

experience in mind, but they exist in practically all instances. Children from rural areas will find certain items relatively easier than do children of equal ability who have been reared in the city, while the reverse will be true for other items. "Only" children whose social contacts within the home have been for the most part confined to adults are likely to be relatively precocious in linguistic development in comparison with other aspects of ability, while twins, especially those of the same sex, who spend a large proportion of their time with each other during which each hears and imitates the other's imperfect speech, are likely to be somewhat retarded in language (Day, 1932; Davis, 1937). All this has a very important bearing on the controversial issue of environmental influence upon mental ability, which will be discussed in a later chapter. At the moment we shall note only that no test has yet been developed in which the difficulty of the items is wholly determined by intrinsic factors. Differences in the external circumstances under which children are reared may be expected to affect their test standing to the extent that extrinsic factors play a part in determining item difficulty. This is equivalent to saying that in theory, at least, it is entirely possible to change a child's test performance without affecting his general behavior or his ability to cope with the situations of everyday life in the least. A sign may be changed without altering the thing signified; a sample may be so biased that it no longer represents the total from which it is drawn. A test which is either a sign or a sample is no exception to these rules. Accordingly a change in test standing does not always or necessarily indicate a change in the trait which the test is designed to measure.

It is important, therefore, when arranging items in order of difficulty, to give careful consideration to the factors which influence difficulty. To the extent that difficulty is intrinsic rather than circumstantial, the order of items may be expected to remain approximately the same from group to group. If the differences in difficulty are largely external in origin, the result of experiences which vary from one social group to another,³ the weights assigned to the various items in the scale cannot hold good for all individuals or groups. This fact makes for inappropriateness of reference standards and instability of the scores earned by the same child at different ages. Although inspection of the items will afford some clues, a comparison of the results obtained for groups known to differ with respect to the factors which seem likely

³ There is, of course, the possibility that success may depend upon experiences which are equally common to all groups for whom the test is designed. But in such cases the item would be eliminated because of its lack of discriminative ability, or, if intrinsic factors are also involved, the latter alone would affect its usefulness in the scale.

to have bearing upon the matter is a preferable method of answering this question.⁴

A number of methods have been devised for equalizing the differences in difficulty within a series of test items. Generally speaking, however, all are based upon the principle stated by Fullerton and Cattell (1892) more than a half century ago to the effect that "differences equally often noticed are equal." Translating this statement into terms of success or failure on the items of a test we have something like this: "Differences in difficulty are equal if the tasks are equally often passed." That is, if 60 per cent of the members of Group A are able to pass Item *a* while only 25 per cent of this group are able to pass Item *b*, and if 60 per cent of Group B (a more able but otherwise similar group) can pass Item *b* but only 25 per cent succeed with Item *c*, then we may say that the difference in the relative difficulty of Items *a* and *b* is equal to that between *b* and *c*. Since it is not convenient in practice to find a sufficient number of items which meet the criteria of item validity described in the last chapter and which also are equally spaced with reference to difficulty, some test makers solve the problem by a statistical conversion of the differences in the percentages who pass each item into relative position on a scale of equal units, disregarding the fact that the spaces from item to item may be of quite unequal length.⁵ Others, particularly in those cases where a large number of items for trial can easily be had,⁶ prefer to try out enough to make it possible to choose a series in which the values are fairly evenly spaced.

A summarized account of the methods commonly used for determining the position of items on a scale of equal units has been given by Guilford (1936). Other textbooks on statistical method also include material on this subject. The point that must be stressed here is that unless the differences in difficulty from one item to the next are determined wholly by intrinsic factors, the values obtained cannot be expected to hold good for groups of markedly different social or educational background from that used in the original derivation of the scale. They

⁴ It is not necessary that the groups compared be closely matched in respect to general ability, since the point in question is not the absolute difficulty of the tasks but only their position with respect to each other, that is, their rank order within each group.

⁵ Such a scale is comparable to a foot rule in which the points marked occur at irregular intervals but their correct measurements are indicated, as 0.4 inch, 1.2 inches, 1.8 inches, 2.7 inches, etc. If the marked points are fairly closely spaced, the loss in accuracy resulting from the use of such a scale, as compared to one in which the marked points occur at regular intervals, is small.

⁶ As, for example, in constructing a spelling scale in which the items may be chosen from those at a predetermined frequency level according to the Thorndike Word List (1921) or some similar criterion.

may differ with such factors as age and sex, even though they were derived from a group including both sexes and with a range of ages corresponding to that for which the test was designed.⁷ They may change with the passage of time as social conditions change. At most they represent an average increase in difficulty, an average that may hold for the group as a whole but does not necessarily hold for all the individuals comprised within the group.

Thus the so-called "absolute scaling methods" cannot be expected to do more than make a partial correction for the irregularities in the spacing of items that usually exist in any series that has been chosen without regard for this factor. The spacing is likely to be more nearly valid if the scale has to do with a comparatively simple and clearly defined trait such as spelling ability; it is likely to be less dependable when the trait to be measured is a complex one and the items used differ from each other in kind as well as in difficulty, like those in most intelligence tests. Even in the latter cases, however, some form of scaling is likely to be distinctly worth while, particularly when the scales are used for group comparison rather than for the appraisal of individuals. When groups are to be compared, however, care should be taken to see that the scale values are sufficiently similar for each to justify the use of a common instrument. If it is found that the values are altered to a significant degree when groups of different composition are used for determining them, that in itself becomes an important scientific fact which merits further study.

THE QUESTION OF THE ZERO POINT

The fact that a given individual is unable to pass any of the items of a given test is, of course, no indication that he is completely lacking in the ability which the test is designed to measure. Of this most people who work with tests are aware, but the correlated fact that test scores cannot justifiably be compared with each other by means of simple arithmetic is frequently overlooked or ignored. It is still unfortunately common to hear teachers say that John did twice as well as Dick on such and such a test, or even that Mary, whose IQ was found to be 50, is only half as bright as the average child. Such comparisons could be justified only if a zero score meant zero ability and if the test items were placed at equally calibrated intervals.

⁷ If, for example, the sexes were combined in the original derivation, any sex differences in scale value that exist would be covered up. The resultant figure would be an average of those pertaining to the two sexes and would not be truly representative of either.

Because it is practically impossible to establish the zero point of a trait by empirical methods,⁸ a number of attempts to do so have been made by statistical means. Perhaps the soundest of these procedures was proposed some years ago by Thurstone (1928). Thurstone reasoned that inasmuch as the variability of the distributions of scores made on tests calibrated according to a scaling method he devised showed a fairly consistent and regular increase with age throughout the developmental period, an extrapolation of the growth curves downward to a point at which variability no longer existed, i.e., to the point at which all scores became alike, would automatically fix the position of the zero point, since in the nature of things variability cannot be negative. By applying this procedure to a number of different tests after reducing the score values to "absolute" (that is, equally calibrated) units, he found that in every case tried the zero point was placed at birth or shortly before birth. This is at least in accordance with reasonable expectation. Although Thurstone's procedure has been criticized by some, it is at least based on perfectly straightforward reasoning and leads to equally straightforward results.

Provided that the limitations introduced by the lack of a clearly defined zero point are clearly understood, the matter is not of as great practical import as might at first be thought. In taking the measurements of a room, for example, it may sometimes be more convenient to start not from a corner but from some other clearly defined point, say a window or a door. If the starting point is known, no misunderstanding need occur. So likewise, if the fact that the zero point of a given test does not coincide with the zero point of the trait which the test is designed to measure is held constantly in mind, with all the limitations of computation and interpretation this fact entails, no problems need arise. The trouble is that many people who are aware of the fact itself so often disregard its necessary consequences.

THE EFFECT OF INSUFFICIENT RANGE OF DIFFICULTY UPON TEST SCORES

That the difficulty of a test should be appropriate to the ability of the subjects with whom it is to be used is self-evident. It would be a

⁸ It would be difficult indeed to demonstrate that any individual who is able to survive is completely lacking in any basic human trait. Even the vegetative idiot who would starve with food well within his reach if it were not put into his mouth, who cannot walk or sit without support, whose glance is wavering and uncertain, who has no means of communicating with his fellows, and who is incapable of learning what seem to be the simplest acts of skill may still be thought to possess the rudiments of intelligence though he certainly may be classed as very near the zero point. But it would be a rash psychologist who would be willing to deny him a minimal degree of the trait.

waste of time to try to use a test of college algebra for appraising the mathematical skill of a group of children just learning the multiplication table. It would be equally foolish to use a reading test designed for children in the first grade in order to measure the reading ability of high school seniors. While no one is likely to do anything so manifestly absurd as these procedures would be, less pronounced instances of the use of tests which are either too easy or too difficult for a considerable number of the group tested are all too common. Inasmuch as

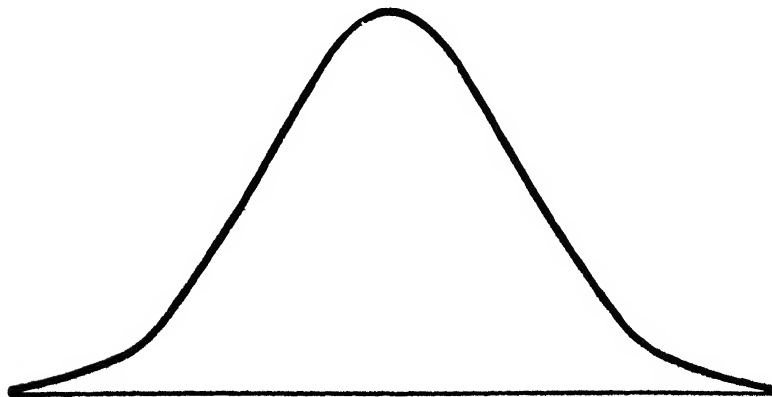


FIG. 4. THE NORMAL PROBABILITY CURVE.

most if not all complex abilities have a qualitative as well as a directly quantitative aspect, two individuals of equal talent (as far as the total is concerned) may nevertheless differ considerably with respect to their ability to handle the parts or items of which that total is made up. If, therefore, a test is to be considered adequate for a given group, it must have a range of difficulty sufficiently great to expose all the deficiencies of even the most backward subjects with respect to the trait it is designed to measure and to permit the more able members of the group to display their talents to the full.

Statisticians frequently designate the lowest level of difficulty within a given test as its "floor," and the highest level as its "ceiling." Now if it so happens that a certain group of children whose actual distribution of ability conforms fairly closely to the normal curve (see Figure 4) are given a test which for them is so difficult that the majority are unable to pass more than a few items of it and the mean score of the group is not far above the floor, their distribution of scores will not be normal but will be skewed in a manner similar to that shown in Figure 5. Although the more able subjects may be correctly measured, the number

of items appropriate to the lower levels is too small to constitute an adequate sample of what the backward subjects can do. Not only will their scores be unstable but they will tend to be too high, inasmuch as they receive full credit for the items with which they are able to succeed

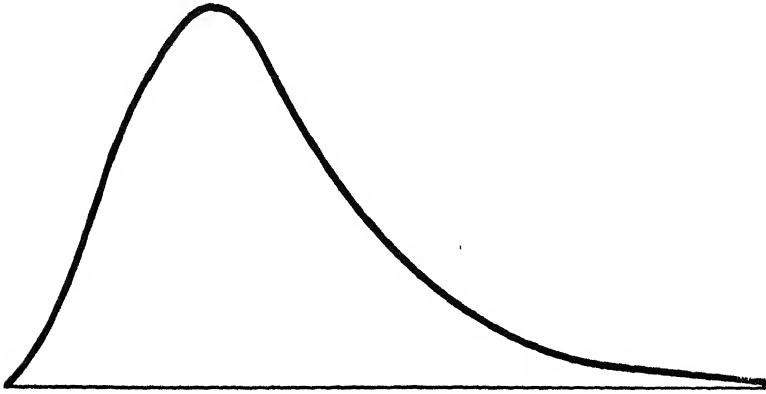


FIG. 5. SKEWED CURVE RESULTING FROM THE USE OF A TEST THAT IS TOO DIFFICULT FOR THE GROUP TO WHOM IT IS GIVEN.

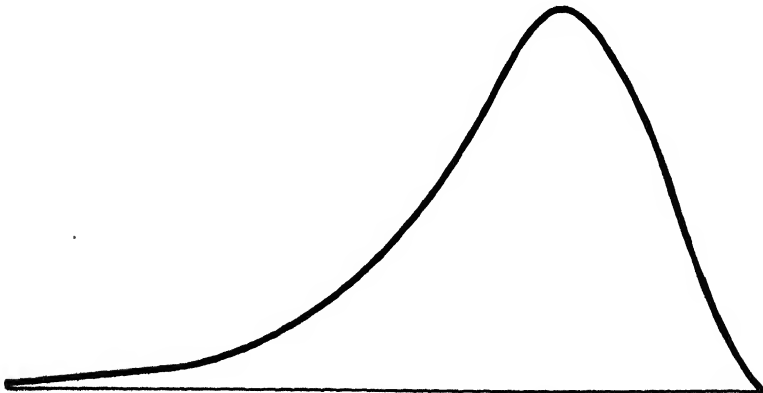


FIG. 6. SKEWED CURVE RESULTING FROM THE USE OF A TEST THAT IS TOO EASY FOR THE GROUP TO WHOM IT IS GIVEN.

and are not penalized for easier items on which they might very possibly have failed had such items been included in the test. In like manner, if a test is too easy, the brighter members of the group will have inadequate opportunity to show what they are capable of doing, and their scores are likely to be too low.

There is good reason for believing that the form of distribution of

most, if not all, mental abilities conforms at least roughly to that of the normal curve. Accordingly a valid test of this kind with a range of difficulty appropriate to that of the subjects to whom it is given will ordinarily yield a score distribution that is relatively free from skewing unless there has been bias in the selection of subjects. Of course if the number of subjects is small, not much can be concluded from the form of the distribution. But if the sample is of fair size, say 100 cases or more, and there is no reason to suspect that bias was present, then such skewing of the distribution curve as appears in Figures 5 and 6 calls for careful scrutiny of the evidence as to the suitability of the test for the group with which it was used.

In an attempt to correct not only skewing but other irregularities in the distribution of test scores, a process known as "normalizing" is often used. It is apparent from an inspection of Figures 5 and 6 that because of the limitations set by the floor or by the ceiling of the test, the differences in ability of the subjects at one or the other extreme are very poorly represented by their scores on the test. Instead of being spread out as they would have been if opportunity had been given to show all the points of superiority possessed by the one group or all the aspects of deficiency of the other, these scores are massed closely together.

Normalizing a set of test scores means assigning such weights to the original scores as will throw them into the form of a normal distribution. There are various ways of doing this, depending on the nature of the data and the assumptions that can justifiably be made concerning it.⁹ All, however, are based upon some method of transforming raw scores into standard scores, a procedure which will be described more completely in Chapter 13. For the present we need call attention only to two points as follows: (1) normalizing usually calls for changing the distance between the scores originally obtained in such a way that at those points where the subjects are closely massed the scores will be thrown more widely apart, while at those points where the subjects are widely spaced the differences between score points will be reduced; (2) although this procedure will assign such positions to the extreme cases as will more nearly indicate the extent of their difference from the majority, no merely statistical method is capable of selecting among the extreme cases those persons whose scores were actually affected by the test limitations or

⁹ Of course if there is reason to believe that the form of the sample distribution is not the result of inadequacy in the test but is a truthful representation of the distribution of ability within the group in question, no adjustment of scores is called for. The best test of such an assumption is to give another test of a similar nature but easier or more difficult as the case may require. If the form of the distribution remains unchanged, the assumption that the skewing was due to the particular selection of subjects and not to the inadequacy of the test appears warranted.

those who would have done no better (or more poorly) regardless of opportunity. No amount of later tinkering with a set of imperfect scores can completely make up for the use of a test which either was badly calibrated in the first place or is unsuited to the abilities of the group to whom it is administered.

CHRONOLOGICAL AGE AS A BASIS FOR SCALING

The use of chronological age as a standard for determining the difficulty of items in a scale designed to measure some function that is known to change with age dates back to 1908, when Binet published the first revision of his original (1905) scale. As originally conceived, the device appeared deceptively simple. The difficulty of a given task would be determined by ascertaining the age which has to be reached before the average child is able to deal with it. But it was soon found that the correct placement of a test item on the basis of age is by no means an easy task. Some items show a very rapid increase in the percentage succeeding with them as age advances; with others the improvement is more gradual. Some items are highly correlated with each other and with the total; in other cases the correlation is much lower. Japsen (1944) has pointed out that because mental age is computed on the basis of all the items passed, only in the unlikely event that its correlation with the others at the same level is perfect would an item be correctly placed at the point where it is passed by half the cases. The lower the correlation between items, the greater must be the percentage succeeding with a given item at its placement level if the average child is to earn a mental age corresponding to his chronological age. That the percentage of cases passing a test item at the age level where it is placed must be greater than the 50 per cent at first thought reasonable has long been recognized. Binet estimated that the correct value lies somewhere between two thirds and three fourths of the cases. Both Terman and Kuhlmann obtained similar results by empirical examination of their data, and the former noted that intercorrelations of the test items as well as the slope of the growth curves for the individual items are involved. However, he believed the problem too complicated to permit a solution that would be practicable for use in the actual placement of items. He noted that if a balance were so maintained that if one item chanced to be placed at too high an age level but another was placed at a corresponding point below its correct position, the errors would cancel each other and the result would be the same as if both had been correctly placed. However, it should be noted that this principle may not hold good in all cases. The usual rule for determining the limits of testing

in the case of an age scale is to begin at the age next below the child's chronological age¹⁰ and continue downward as necessary until an age group is reached at which all the items are passed. This is known as the basal age. In like manner the examiner then proceeds to try the items at the ages above the basal year until an age group is reached at which all are failed. The mental age is then obtained by adding to the basal year a proportional number of months' credit for each test item passed above the basal. If, as is true with the 1937 Stanford-Binet, six items have been included in each yearly age group,¹¹ then each item counts for one sixth of a year's growth or two mental months.

Now if it so happens that the misplaced tests occur at one of the limits at which the testing is discontinued but not at the other, the balancing will fail to work, and the obtained mental age will be either too high or too low, as the case may be. Moreover, the time required for testing will have been unnecessarily increased by the necessity of administering an additional year's series of tests which, except for the misplaced item, would have fallen outside the required range. Thus the correct placement of items in a year scale is a matter of greater importance than has sometimes been supposed, not only because of the facts just mentioned but also because incorrect placement of items may result in irregularities in group variability which seriously affect the meaning of the intelligence quotient. This point will be discussed more completely in Chapter 11.

MENTAL-GROWTH CURVES

That the curve of mental growth when plotted against time is not a straight line is conceded by practically all authorities, but there is no general agreement as to the best method of fitting the curves, nor is there uniformity in the results obtained by different methods of curve fitting or even in the results obtained by the same method when applied to the data from different tests or from different populations. That the greater number of mental functions undergo their most rapid changes during the early years of life seems well-nigh certain. However, intelligence does not share in the growth spurt preceding adolescence which is

¹⁰ Unless, of course, other facts are known in advance which justify the expectation that the child is either much accelerated or retarded in his mental development, in which case an appropriate adjustment will be made in the point at which the testing is started.

¹¹ Except at the early ages, where the items are arranged in half-year groups and each item counts for one mental month instead of two. At the upper extreme, also, the age groups are differently spaced and the number of months' credit for each item differs accordingly.

so characteristic of physical growth.¹² Taken together, these facts point to the assumption that the most probable form of the curve of intellectual growth is a parabola of which the exact formula has not been determined. Quite possibly it may differ for various population groups and almost certainly it will alter to an unknown extent with test content. In any case the straight line assumed by the age-scale method of standardization ought not to be regarded as a growth curve at all but merely as a time scale along which samples of equal size have been arbitrarily drawn to represent the various levels of attainment at successive ages.

Whether or not there is warrant for assuming a single form for the growth curve of such a complex trait as intelligence is a moot question and one to which there can as yet be no unqualified answer. That some degree of uniformity is apparent in the results obtained by different students of the problem in spite of diverse conditions and populations suggests the possibility of a basic pattern of growth which is but imperfectly represented by the instruments and procedures at present known to us, but this possibility requires much further verification before it can be accepted as fact.

A special aspect of the mental-growth curve that has attracted some general interest is the location of the midpoint. At what age has the average person attained one half of his ultimate mental stature? Thorndike (1926) drew up a growth curve based on the results of his CAVD test, which had been extended downward to a point as close to zero intelligence as tests could be devised. When this scale was applied to school children, infants, and low-grade feebleminded children in institutions,¹³ a curve of growth plotted in units of equal difficulty ("altitude" in Thorndike's terminology) was drawn up. According to this curve, the midpoint of mental growth is reached just before the third birthday. If this sounds fantastic to the reader, he is invited first to compare the abilities of a newborn baby and an average adult, and then, by slowly moving up the age scale, to try to locate the point at which the differences between the two levels are equally balanced, that is, the point at which he is no longer able to say with any feeling of certainty that at the age

¹² Although children of the same chronological age have been found to differ rather markedly in nearly all physical measurements according to whether they have or have not attained puberty, and although a good many tests of interests and attitudes also yield reliably different scores for the two groups, none of the standard tests of intelligence have been found to show any consistent relationship to physiological maturity. Even the comparatively rare cases of markedly precocious puberty in which sexual maturity may be reached as early as the age of two or three years are no exception to this rule.

¹³ The lower end of the curve was based on a study by K. S. Cunningham, a student of Thorndike's whose report appeared a year later (1927).

in question the child's abilities are more nearly like those of the newborn or those of the adult. If this is done with due regard to such matters as language, locomotion, self-help, and evidences of remembering, reasoning, and so on, Thorndike's figure may not seem unreasonable.

In summary, then, we note that even after a series of suitable test items has been selected, their organization into a scale is not a simple matter. Questions of their relative difficulty expressed in units presumed to be equal, of their intercorrelation with each other and with other criteria, of the location of the zero point, and similar problems continue to vex both the test maker and the test user. There is also the question of the choice between the "point-scale" method of organization in which the items are so chosen as to fall at approximately equal intervals on a scale of difficulty, like the successive inches on a foot rule,¹⁴ or the "age-scale" method in which the items are placed in groups according to the age at which they can be passed by such a percentage of cases as will make the average child earn a mental age corresponding to his chronological age. Each of these methods has certain advantages and each presents its own problems. The age-scale method has always been more popular with teachers and other practical workers since the mental-age concept provides them at once with a workable idea of the level of performance to be expected of each child. This information has been found so valuable that point scales are usually translated immediately into mental ages as a means of interpreting their significance. Research workers are likely to prefer the point-scale method if the work of calibration has been carefully done. To date, however, few point scales have been as adequately standardized as was the 1937 revision of the Stanford-Binet, which is still recognized as the nearest approach to a basic standard in the field of intelligence testing that is available at the present time.

The question is sometimes raised as to the desirability of setting up a series of uniform standards in the form of mental tests which would be comparable to those maintained for physical measurements by the United States Bureau of Standards. These tests would then serve as criteria with which other tests could be compared. Theoretically such a plan would simplify many problems; practically it is doubtful whether it would be either a feasible or a desirable one. Changing social conditions are likely to alter the form in which mental traits are manifested, even though the characteristics themselves may remain fundamentally the same. As a result, some test items tend to become obsolete after a lapse of time although many of them may be expected to maintain their value

¹⁴ Or their relative positions on such a scale are indicated, as noted on p. 144.

for long periods. Moreover, improved skill in test construction has caused new and presumably better methods to supplant the earlier and less perfect ones. This process of improvement is by no means complete. Thus, even if it were feasible to establish national standards at the present time, it would certainly be undesirable to do so, for to give lasting status to an unperfected standard would stultify progress.

PSYCHOLOGICAL MEASURES COMPARED WITH PHYSICAL MEASURES

Even after every effort has been made to equalize the differences from one step to another on a scale of psychological values, the units of measurement thus derived are still relative rather than absolute.¹⁵ In the first place they relate to a particular scale only. They cannot safely be transferred to another scale designed for the same purpose. Moreover, since all units of measurement are derived from the performance of the particular sampling of subjects used for standardization, they may or may not retain the same position with reference to each other when the scale is applied to other groups who differ from the reference sample in respect to such factors as age, sex, or experiential background. For example, Cunningham (1927) found that even with so carefully standardized a test as Thorndike's CAVD (1926), differences between the performance of adult imbeciles and that of normal children between the ages of two and a half and six years on many of the separate items of the scale were so great that they would be unlikely to occur by chance more often than once in many thousand trials. It is evident that the approximate equality of the successive steps on this scale which had been obtained for the standardization group did not hold good for groups of such markedly different composition as those studied by Cunningham.

Even for the group from which they are originally derived, scale values may change as experience changes. The effect of previous experience on test scores has been the subject of many investigations, but for the most part these studies have dealt only with total scores and have given relatively little attention to the possibility of changes in the scale values assigned to the separate items. It has been found, however, that certain items are much more subject to change with practice than others are. This difference in "practice effect" almost of necessity results in some modification of the scale values assigned to the different items when a test is repeated.

¹⁵ The term "absolute" as applied to scale values means "characteristic of an object or phenomenon by itself as distinguished from its relations to other objects or phenomena (except the standard)."—Warren (1934).

Not only previous experience on the part of the subjects but knowledge of the previous performance of tested individuals on the part of the examiner may result in unconscious but significant changes in the administration and scoring of tests. The psychometrist who has been indoctrinated with the belief that the IQ is always "constant" and who is aware that a particular child made a poor showing on an earlier occasion is likely to be far more willing to accept his "I don't know" as valid or to score a marginal response as a failure than he would if the child's earlier standing had been high. Since some test items give greater leeway for variation in procedure than others,¹⁶ it is altogether likely that scale values as well as total scores are subject to the halo effect arising from an examiner's predisposition to expect certain children to do well and others to do poorly.

Psychological measures, therefore, differ from physical measures in a number of ways. Since there is no generally accepted series of standards comparable to those established by the United States Bureau of Standards for the uniform calibration of physical instruments of measurement, each test maker is obliged to do his own calibrating in terms of his own data. Although the gradual acceptance of certain assumptions and the more or less uniform statistical methods arising out of these assumptions have brought some degree of order into what otherwise would have been well-nigh hopeless confusion, it must not be forgotten that such terms as "absolute scaling," "equal units," and the like are more hopeful than accurate. At most they apply only to a particular group of subjects at a particular time. Physical measures do not change with mere repetition; psychological measures inevitably provide some opportunity for learning,¹⁷ and thus the standards used for a first test usually require some

¹⁶ In the Stanford-Binet, for example, the instructions for some items may be repeated at the examiner's discretion; for others no repetition is permitted. The amount of urging or praise to be used may also be varied, in most cases, in accordance with the examiner's judgment of what is required in order to ensure maximal cooperation from each child. There are also many intangible factors, hard to describe or control, which nevertheless have a definite effect upon the attitudes and behavior of children in the test situation. Someone has well said that children respond to the "muscle tensions" of an adult quite as markedly as to their words. Slight changes in the examiner's posture, little impatient movements of his hands, a lifted eyebrow, a smile, or a frown may be enough to cause a timid child either to give up a task as hopeless or to continue working with renewed zest. Bright children in particular are quick to profit by these expressive movements of the examiner, which they utilize as cues for guiding their responses.

¹⁷ The effect of previous practice upon mental test scores varies with a number of factors. Certain kinds of test, notably form-board tests which are so extensively used in many of the so-called "performance" (nonverbal) scales, are so greatly affected by practice as to render the standards used for a first trial largely or wholly invalid when the test is repeated, even though a considerable period of time has elapsed between the testings. At the Institute of Child Welfare of the University of Minnesota

modification if the test is repeated. Even when the test is entirely new to them, children who have frequently been tested are likely to show some adaptation to the test situation as such which enables them to deal with it more effectively. College students who have become accustomed to examinations of the so-called "objective" type are likely to do somewhat better on tests of this kind than others of equal ability whose experience has been largely or wholly confined to tests of the "essay" sort. Something is gained just by learning the trick.

we have found that most children recognize tests of this kind and make a disproportionately better performance on them than could be accounted for on the basis of mental growth alone, even after a year or more has intervened. In tests of the Binet type, some items are more affected by practice than others. Bright children may, and often do, remember some of the items about which they felt uncertain at the time of testing. By questioning parents or teachers they ascertain what the correct answer should be and are ready with the official response if a retest is given. There may even be some passing on of information of this kind among the children in a school where much testing is being carried out. In addition to such informal and unintentional results of test experience, there is, of course, the possibility of direct coaching on the test items. Many of the items in the more widely used intelligence scales have become matters of common knowledge among well-read people, especially those whose college training has included some work in psychology. Interest in the mental development of their children not infrequently leads such parents to "try out" these items with them and in some instances to drill them on the answers in a well-meant attempt at child training or even with the hope of creating a child prodigy. While such coaching is not common, it occurs often enough to make it essential for the psychometrist to be on guard for it. Here the fact that some test items are much more subject to the effect of coaching than others is a matter of prime importance to the examiner, for it means that the child who has been coached is likely to show a very different pattern of responses from that of children in general, and the examiner who is alert to the qualitative as well as the grossly quantitative aspects of the child's test performance can usually detect such cases without too great difficulty.

Age Standards and Quotients as Interpretative Measures

THE NEED FOR UNIFORMITY OF QUANTITATIVE TERMINOLOGY

In Chapter 10 the advantages of establishing some kind of uniform system of expressing the results of mental measurements were pointed out. In the absence of common linguistic terms, it becomes practically impossible to interpret measurements of the different characteristics of a given person. Even more serious is the fact that without the aid of symbols with constant meaning it becomes very difficult to make much progress in quantitative method. This is as true with respect to the practical aspects of mental testing as it is in the more scientific area of test construction. The lack of terms which can be understood by those who are directly responsible for the welfare of children or for the guidance of older individuals would render it difficult, if not impossible, for them either to make use of test results in their own work or to provide scientific workers with the kind of direct evidence of test validity essential for effective progress in test construction and in the improvement of existing methods.

But uniformity may be assumed when it does not exist, and it is quite as important for the practical worker to know and understand the factors which make for dissimilarity in the meaning of a presumably constant term as it is for him to know its theoretical applicability to a concrete situation. It is unfortunately true that understanding of the limitations of a given method so often lags behind information as to its possible usefulness. This lack of understanding often leads to mistaken methods of handling individual cases in practical situations or to mistaken theories and conclusions on the part of research workers. Most of the terms used for the interpretation of test results have definite limitations of meaning which do not interfere with their value if these

restrictions are understood and respected but which, if ignored, may have highly unfortunate consequences. Although these sources of error have frequently been pointed out by test makers and statisticians, they are too often overlooked, not only by educators, social workers, and others with little training in scientific method, but also by many research workers whose desire to solve problems runs ahead of their methodological knowledge and understanding. It therefore seems worth while to devote some time to a consideration of the statistical assumptions underlying the development of some of the methods commonly used for the interpretation of mental tests and of the statistical limitations and possibilities inherent in their derivation. Because age and quotient methods have the longest history and have been more extensively used than any of the other interpretative devices, these will be considered first.

MENTAL AGE AS AN INTERPRETATIVE MEASURE

Few if any psychological concepts have had such a tremendous influence upon the practical handling of children as Binet's introduction of the idea of mental age as a criterion by which their ability to learn and the level of behavior that may fairly be expected from them can be gauged. The great advantage of mental age as an interpretative measure lies in its concreteness. Even persons of little education and no psychological training whatever usually have some idea, crude though it may be, of the differences in ability that separate the child of ten from the child of seven or the six-year-old from the two-year-old. It thus becomes comparatively easy to help the teacher or parent to realize that twelve-year-old Sam with a mental age of six can be expected neither to do the academic work of the sixth grade nor expected to profit by the methods of teaching appropriate to the twelve-year-old level of understanding. In like manner many problems of parent-child relationship may be lessened if parents can be brought to see that eight-year-old Sally is mentally as capable of managing herself and her own affairs as is the average twelve-year-old and should be treated accordingly.

The scientist, however, will realize that statements such as the above are greatly oversimplified. The twelve-year-old may have done no better than the average eight-year-old on some particular test, but it must not be forgotten that another test might have yielded a different result, though it is unlikely that it would have classed him as a child of normal ability.¹ Even if we ignore the obvious differences in physical

¹ Assuming, of course, that the test used was one of recognized merit and that it was properly administered and scored. Not all tests are equally reliable or valid; not all examiners are equally competent.

size and physiological development that are likely to exist, even if we also disregard such factors as life experience and length of schooling, we cannot afford to overlook the fact that no test provides more than a very limited sample of the universe designated by its name. Moreover,

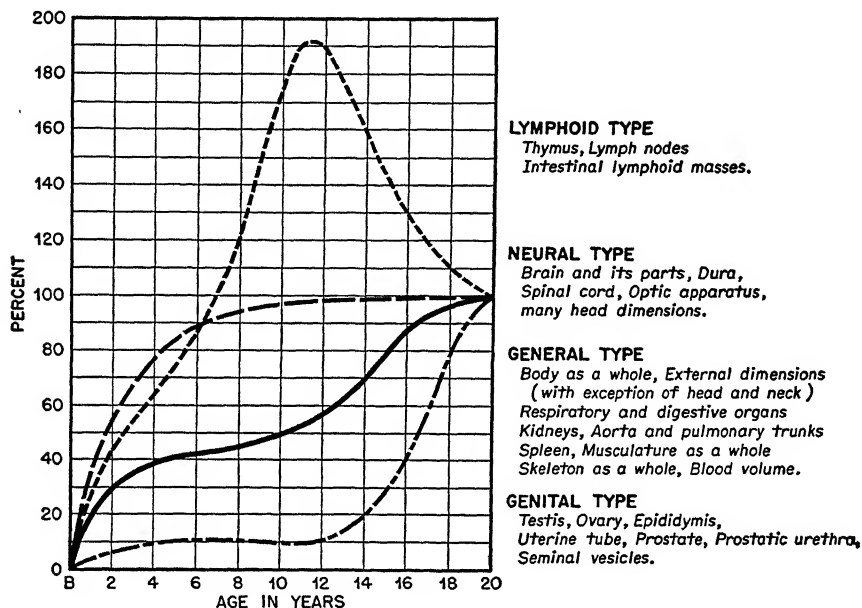


FIG. 7. FOUR MAIN TYPES OF POST-NATAL GROWTH. (Reproduced by permission of the publishers from "The Measurement of the Body in Childhood" by R. E. Scammon in *The Measurement of Man* by J. A. Harris, C. M. Jackson, D. G. Paterson, and R. E. Scammon. Minneapolis, Minnesota: University of Minnesota Press, 1930.)

because the reference standards for different tests have usually been derived from different and sometimes quite diverse samplings of subjects, the mental age earned on one test may differ considerably from that which would have been obtained had a different test been used, even if both were completely freed from errors of measurement² in the statistical sense of the term.

Finally it must be remembered that the spacing of items in a mental-

² As statistically defined, an error of measurement refers only to changes on successive administrations of the same scale or of different forms of a scale which are presumed to be wholly comparable to each other when the testings take place over so short a time interval that no significant growth change in the subjects tested can have occurred. A specified allowance for practice effect is not counted except as this allowance may prove erroneous. Errors of measurement as thus defined have no reference whatever to the purpose for which a test was designed but only to the stability of the results obtained by its use.

age scale is a function of time and not of growth. The distinction is readily seen if one compares mental growth with physical growth. The average gain in height made by a group of normal children differs considerably from one age period to another. (See Figure 7.) Although it is unsafe to reason directly from structure to function, it nevertheless seems very likely that the great precocity of neural growth as compared with growth in other types of body structure may be mirrored, at least to some extent, in the growth of mental ability. It will be recalled that Thorndike (1926), in his attempt to project his CAVD test downward to the level of early childhood and infancy, found that the midpoint of growth in "mental altitude," as he termed the ability to perform tasks of increasing difficulty, is reached at about the age of three years or slightly earlier than that. This may seem a startling idea to those who are familiar with the fact that the average child continues to develop mentally until around the age of sixteen years.³ But it must not be forgotten that growth in intelligence, like growth in height, must be measured in terms of its own units and not in terms of the length of

³ The exact age at which mental growth ceases is still a matter of controversy. Some have placed it as low as fourteen or even thirteen years. The latter figure is based chiefly upon the results obtained from the use of the Army Alpha tests during World War I, where it was found that the average score made by the men in the draft army corresponded fairly closely to that made by school children at the age of thirteen. There are two sources of error, however, in such a comparison. First, there is the factor of selection. The draft army did not constitute a representative selection of the total male population at that age since it did not include those in the officers' training camps or many of those who volunteered without waiting to be drafted, nor did it include the medical corps. In the second place, the comparison is made along the wrong regression line. To say that the average score made by the soldiers was no better than that made by the average child of thirteen is not the same thing as ascertaining the chronological age at which scores on the test in question cease to improve. It may also be noted that in many instances the physical conditions under which the tests were given were far from ideal. In a test which emphasizes speed as greatly as does the Alpha, poor lighting or the absence of suitable writing conditions may have an extremely detrimental effect upon scores.

The fact that growth does not stop abruptly but slows off by gradual and well-nigh imperceptible degrees and the further fact that with advancing age it becomes increasingly difficult to distinguish between the results of learning and the capacity to learn are additional obstacles in the way of establishing the precise age at which mental growth ceases. A still further complication arises from the fact that there are, in all probability, individual differences in this respect. Not all persons attain mental maturity at exactly the same age, any more than is true with regard to growth in height.

The weight of modern opinion, however, seems to indicate that some small increment in intellectual growth continues at least up to the age of eighteen years in the majority of cases, but the absolute gain after the age of thirteen or fourteen is very small. The curve probably begins to flatten out during the preschool period. According to Thorndike (1926), by the age of 6½ years about 83 per cent of the ultimate level of mental altitude as measured on his CAVD test has already been attained.

time required to attain a given mental stature. There is no reason whatever for assuming that the average gain accomplished from one age to another is equal, for although growth takes place in time, there is no more reason for assuming that it conforms to time in a quantitative sense during the period while it is continuing than during the later period when mental growth has ceased but time continues to advance.

Mental age, then, is not an absolute measure of growth but only a convenient means of interpreting growth. If properly understood it is perhaps the most valuable aid to the understanding of a child's mental capacity that has ever been devised, but its major usefulness lies along practical rather than scientific lines. It is true that in the course of developing mental-age standards much has been learned about the qualitative aspects of mental growth. Although this knowledge constitutes an important part of our scientific information about child behavior and development, the use of mental age for the solution of scientific problems is restricted to those investigations in which the mathematical treatment of the data involves no assumptions as to the equality of the growth units employed. This is equivalent to saying that mental ages should not be added, subtracted, or averaged.⁴ They are interpretative measures only.

THE INTELLIGENCE QUOTIENT

As was previously noted, the use of the ratio of mental age to chronological age, now known as the intelligence quotient, was first proposed by Stern and also by Kuhlmann as early as 1912, but it was not until Terman made it an integral part of his 1916 revision that it attained the widespread popularity it holds today. By that time the need for a more generalized way of expressing the extent of backwardness or acceleration shown by an individual child had become acutely felt. It was obvious to all who worked in the field of mental tests, and especially to those who were primarily concerned with the use of tests in selecting children for placement in institutions for the mentally defective or in special classes in the public schools, that a simple statement of the number of years of retardation without regard to the length

⁴ A few exceptions to this rule may be made. If a child is given the same test twice with so short a time interval between the testings that no appreciable mental gain is likely to have occurred, any difference in the results may safely be attributed to errors of measurement rather than to differences in the step intervals between successive mental ages. Therefore the average of the two may be taken as a better indication of his true mental level than either one separately. If two forms of the same test are available and both are given, the same principle may apply, provided that the equality of the two forms at each age level has been carefully established.

of time during which this retardation had been accumulated could not give an accurate picture of the facts. The improbability that a truly feeble-minded person can ever be made normal had been recognized since the days of Esquirol, but the idea that mental tests may also be used for predicting the later mental progress of normal and superior children was something toward which psychologists were then only just beginning to grope.

The 1916 revision was the outcome of several years of work on the part of Terman and his assistants. During this time many children of all levels of intelligence had been retested, some of them repeatedly, after varying intervals of time. What chiefly impressed Terman with respect to these retests was the amazing consistency with which the ratio of mental age to chronological age was maintained in practically all cases. It was this fact, chiefly, which led him to abandon the practice of expressing a child's mental level in terms of the difference between his mental and chronological ages in favor of the ratio between them. To this ratio (mental age/chronological age) he gave the name "Intelligence Quotient." He suggested that in writing the results, the initial letters (IQ) should be used without the abbreviation sign and that the decimal point should also be omitted. Thus, instead of recording that John Smith was found to have an I.Q. of 1.25 one would express it more compactly as an IQ of 125.

Although Terman's belief in the relative consistency of mental-growth rate was strongly implied throughout his discussion of the new scale, in *The measurement of intelligence* (1916), he was very cautious in making absolute or dogmatic statements with respect to prediction. Only once, on page 68, did he refer to the "constancy of the IQ" and then only tentatively as follows:

The inference [from data on the similarity of the IQ distributions at various ages] is that a child's IQ, as measured by this scale, remains relatively constant. Retests of the same children, at intervals of two to five years support this inference. Children of superior intelligence do not seem to deteriorate as they get older, nor dull children to develop average intelligence. Knowing a child's IQ, we can predict with a fair degree of accuracy the course of his later development.

During the next three years Terman and others who were using the new scale set about testing the theory of IQ constancy with vigor and enthusiasm. Many retests were given to children of all ages and social levels, and the effect upon the consistency of the results of such factors as age at testing, interval between tests, sex, social class, and home environment was studied. In *The intelligence of school children* (1919)

the results of a number of these studies are compared and summarized. The findings are presented in the form of graphs, tables, and scatter diagrams, together with a number of reports of individual cases. A notable feature of this report is a diagram showing the correlation between the IQ's obtained on first test and retest for 428 cases of various ages who had been retested after intervals ranging from a few days up to seven years. The correlation (Pearsonian r) was found to be $+.933$. This comparatively high figure is largely due to the wide dispersion of the IQ's, the effect of which in raising the value of the correlation coefficient was not well understood at that time. The S.D.⁵ of the IQ distribution for the first test is 21.8 points, a figure considerably higher than that found for most populations where no selective factor has been at work. The reason for this high variability is to be found in the special interest felt at that time in studying the intellectual extremes—children thought to be possibly feeble-minded on the one hand, and those displaying unusual precocity on the other. The inclusion of a disproportionately large number of these cases in the group tested would, of course, greatly increase the dispersion of the scores made.

A summary of the findings from a large number of different authors leads to the conclusion that the variability of the IQ's of a representative

⁵ For a discussion of the standard deviation (S.D., or as it is often written, σ) the reader is referred to any of the standard textbooks on statistical method. Briefly, the S.D. is a measure of the extent of the individual differences within the group in question, that is, a measure of the group's variability. It is found by taking the difference between the group mean and each individual score, squaring these differences, summing them, and dividing by the number of cases to find the mean of the squares. The square root of this mean is the standard deviation. In practice, certain short cuts which reduce the labor of computation are commonly employed.

There are two reasons why the S.D. is commonly preferred to the more simply obtained *average deviation* (A.D.), which is the mean of the differences between the individual scores and the group mean. In the first place, the S.D. is a more stable measure than the A.D. It varies less from one sample of a given universe to another. Second, the S.D. enters into so many other statistical formulas in common use that, on the whole, the additional time required for its computation will be more than made up by the possibility of using it in other calculations which are likely to be needed.

The product-moment coefficient of correlation, for which the conventional symbol is r , is an example. The formula for r is as follows:

$$r = \frac{\Sigma xy}{N(S.D._x S.D._y)}$$

in which the Greek letter Σ means "the sum of" and the letters x and y denote the scores made by the individual subjects on each of the two measures correlated when each is expressed in standard units, i.e., in terms of the number of standard deviations by which it exceeds or falls below the mean of all the cases included. The numerator of the fraction is the sum of the products of these standard scores. The denominator is the product of the standard deviations of the two correlated measures taken as many times as there are cases.

sample of American urban school children as obtained by the 1916 revision is approximately 16-17 IQ points. It is possible to determine statistically what the correlation between test and retest for Terman's group would most probably have been if a sufficient number of the extreme cases had been removed to make the sample correspond more closely to that of an unselected population. Taking 16.5 as the S.D. of such a group, the most probable correlation becomes $+.812$,⁶ which corresponds fairly closely to what has generally been found for this test by other workers using more nearly typical groups of subjects.⁷

⁶ See pp. 415-419 of *Psychometric methods* by J. P. Guilford (1936).

⁷ There are a number of different ways of gaining a more concrete understanding of the meaning of a coefficient of correlation. Since the extent of agreement between two measurements, either of the same or of different variables, is so frequently expressed in this form it is well for the student to keep certain facts in mind. In the first place, a correlation is not a percentage but it can be reduced to percentage form in terms of what is called the *index of forecasting ability* for which the formula is

$$1 - \sqrt{1 - r^2}.$$

By the use of this index it becomes possible to ascertain the average percentage of improvement over sheer chance which is gained by using the scores on one measurement as a means of predicting those to be made on another. But since the magnitude of this percentage is directly dependent upon the size of r , it is not a fixed quantity but will change as r changes. In the example just given, for instance, if we take Terman's originally reported r of $+.933$, the index of forecasting ability is .64 or 64 per cent better than chance; but if we remove enough of the extreme cases to reduce the S.D. to 16.5, thereby making the group more nearly similar to that likely to be found in an ordinary school system with respect to the proportion of very bright and very backward children included, the r is reduced to $+.81$, which permits an improvement of only 42 per cent over that to be had by chance when the first test is used as a means of predicting IQ's on the second.

The apparent inconsistency in these results is readily understood if we consider a more obvious example. You who read this have a certain ability to estimate the height of other people. Although this ability may change with experience and training, at any given time it may be regarded as a fairly constant skill. If, then, you are set the comparatively easy task of estimating the heights of each of a group of persons whose actual measurements vary from that of a newborn baby (about 19 inches) to that of an exceptionally tall man (about 76 inches) with the others scattered fairly regularly in between you will make some errors, to be sure, but in comparison with the range of differences within the group, these errors will not seem very important. There is practically no likelihood that you would estimate the infant's height as greater than that of the man 76 inches tall and there is not likely to be confusion even when the measured differences are much smaller than this. The correlation between your estimates of height and the actual measurements will therefore be high, perhaps as high as $+.9$. But if your subjects had all been of nearly equal height, let us say with a range of only 62 to 63 inches, a very small error on your part will have serious consequences, as far as the maintenance of the correct rank order of the subjects is concerned. The correlation between your estimates and actual height is therefore almost certain to be low; perhaps it may be zero. Nevertheless your ability has not changed. You are exactly as good a judge in the one case as in the other; the difference lies in what was required from you. In the one case, only very coarse

A more direct and easily interpreted way of determining the constancy of the IQ is in terms of the distribution of changes in test standing actually found when the same children are retested after a period of time. The study by Terman already mentioned presents these facts for the 428 cases used in computing the reported correlation of $+ .933$. The figures which he cites have so frequently been misquoted that it may be well to repeat them here.

<i>Deviations as great as or greater than</i>	<i>Actual frequency (per cent)</i>
5	50.0
10	16.6
15	6.2
20 or more	1.85

It is apparent from these figures that even at this comparatively early period in the history of testing, the fact that large variations in test standing sometimes occur was well known. Taking the figures given above as a basis for computation, we find that in an average school with an enrollment of five hundred children, we should expect to find changes in IQ as great as 20 points in about nine or ten of the subjects upon retest, and changes as great as 15 points in thirty or more. Terman's figures are based upon the results obtained by well-trained and experienced examiners and for this reason less variation is to be expected than would be the case had the tests been given by student examiners or others with little experience. It is clear, however, that the term "constancy of the IQ" is merely an expression of probability and by no means a guarantee of certainty. It is thus not always necessary to invoke special causes for IQ change in individual cases. Since the days of Gauss and Laplace it has been recognized that in a chance distribution of errors a few will accumulate to give large totals at the extremes, and that the cases in which these accumulated changes appear do not of necessity differ in any basic way from those nearer the midpoint of the distribu-

distinctions were necessary; in the other, such fine discriminations were called for that only a relatively delicate instrument could perform them satisfactorily.

Few things in the field of mental testing have caused more confusion than the practice, still all too common, of reporting correlations without regard to the variability of the groups from which these correlations were derived. One still hears many people who should know better say of a given test, "It has a reliability [self-correlation] of $+ .92$," or, "Its correlation with Stanford-Binet is $+ .76$." Such statements mean little. A correlation coefficient merely tells how well the instrument in question has performed a particular task but it does not tell how difficult that task was. Without this additional information the correlation cannot be interpreted. It is like saying that a child read a page without error, but unless we know whether the page was taken from one of Emerson's essays or from the story of *The little red hen* in a primer we can have little idea of his skill in reading.

tion.⁸ If 1000 samples of 100 throws each are taken, using an absolutely unbiased coin and making the tosses in such manner as to ensure that the results will not be prejudiced in one direction or the other, not all the samples will show an equal proportion of heads and tails. The distribution of the expected proportions will follow the binomial theorem and will therefore include some samples that depart quite materially from the 50-50 mark, but this means only that chance is operative. Every test that has ever been developed is to some extent affected by such chance factors as momentary shifts in attention or effort, the physical or emotional condition of the subject at the time of testing, recency of experience with material similar to the test items, and a host of other conditions which may operate either to raise or to depress the obtained IQ beyond its "true" value.⁹ If it so happens that the factors making for a negative change preponderate over those making for a positive change in some individual cases (as sampling theory leads us to expect), these persons will show a loss in IQ upon retest, but it does not necessarily follow that all such cases can be attributed to the same distribution of antecedent factors. This fact is so obvious that it would seem needless to emphasize it were it not for the repeated attempts on the part of zealous but poorly trained persons to "prove" some biological or sociological theory by misinterpretation of the laws of probability. Studies of IQ changes may lead to highly important conclusions if

⁸ That some special factor may have been responsible for change is, of course, also possible. For example, Rust (1931) showed that resistance during the test situation may lower a child's IQ as much as 25-35 points below that obtained after better cooperation was secured.

⁹ The statistical definition of a "true score" is "the average of an infinite number of comparable scores" (Kelley, 1923). Since it is obviously impossible to make an infinite number of measurements of any human being, a statistical approximation which is based on (a) his actual score on a single measurement, (b) the variability (S.D.) of the group to which he belongs, and (c) the correlation between first test and retest of this group or of another which is similar to it in all essential respects is all that can be had. The formula is

$$\text{Estimated true score} = r_{11}\bar{X}_1 + (1 - r_{11})M$$

where r_{11} is the self-correlation of the test in question;
 \bar{X} is the child's obtained score or IQ on this test; and
 M is the average score for his age (100 if IQ's are used).

The accuracy of this approximation is indicated by the formula:

$$S.D. \text{ est. true score} = S.D. \sqrt{r_{11} - r_{11}^2} \quad (\text{Kelley, 1923})$$

which indicates the most probable distribution of the differences between the true scores and the obtained scores of a representative sampling of subjects drawn from the same population as that from which the r and the $S.D.$ used in applying the formulas were derived or one closely similar to it.

appropriate statistical methods are employed by those having sufficiently close acquaintance with the actual testing situation to enable them to set up hypotheses and devise crucial ways of testing them. But the pitfalls in such experiments are many, and few indeed are the experimenters who have succeeded in dodging all of them. The difficulties inherent in what at first was thought to be a relatively simple problem are attested by the fact that after more than thirty years have passed and many hundred investigations have been carried out, the question of IQ constancy is still as live an issue as it was in 1916. It is true that the original question: Does the IQ tend to be constant for the individual? has been broken up into a number of more specific questions such as: How do such factors as age, sex, family background, the particular test used, education, special coaching, and the like, either singly or in combination, affect IQ constancy? And more particularly: To what extent and under what circumstances is it possible to change the IQ and how can such changes be brought about? Difficult as these questions are, it must not be forgotten that the real question upon which their significance largely depends is infinitely harder: What conclusions are we justified in making from such changes if they occur?

A test is a sample which may be either representative or biased. If representative, the sample may be looked upon as a valid sign from which the nature of the total may be inferred with a determinable degree of probability. If the sample is biased, the sign becomes correspondingly misleading. *It is far easier to change a sample than it is to change the universe from which the sample was drawn.* This is not to say that a change in the IQ of an individual child is not meaningful. If the universe represented by his ability to perform tasks of an intellectual nature has changed and the retest is a valid sample of that altered universe, an IQ change is significant. But an IQ change alone is insufficient to show that such a universal change has taken place. For example, as early as 1928 Greene demonstrated that coaching either on the actual items of the Stanford-Binet on which they had previously failed or on other material similar to but not identical with these items brought about a substantial increase in the test standing of school children. Some improvement was still apparent for as long as three years after the original training, even though much of its effect had necessarily disappeared as a result of the fact that a large part of the later testing was on a portion of the scale not covered in the original testing, and consequently not included in the coaching. Other studies have confirmed Greene's finding. It would be a rash psychologist who would contend that a change in general intelligence at all commensurate with the change in test score resulting from the coaching had taken place.

FACTORS AFFECTING THE MEANING OF AN INTELLIGENCE QUOTIENT

The early assumption that an IQ has the same significance at all ages and under all ordinary circumstances, provided that the established rules of testing have been fulfilled, has been shown to be erroneous. Age at testing, the interval between test and retest, the sex of the subject, the character of the test used, as well as a number of statistical factors with which many users of tests are unacquainted and others too often overlook, affect the IQ in various ways and consequently have a bearing upon the conclusions which may justifiably be drawn from it with respect to the intellectual ability of a given subject. Only a few of these factors can be mentioned here. As has previously been pointed out, tests for young children are very undependable indicators of their later intellectual status. This unreliability is quite to be expected on the basis of sampling theory, since a test can only sample the abilities and skills that have already developed and there is little overlap between the universe of intellectual skills that are present in childhood and those by which the varying abilities of adults can be differentiated. Both Bradway (1945) and Maurer (1946) have shown that the criteria commonly used for judging the usefulness of test items for young children at the time of standardization afford little basis for estimating their value in predicting intellectual status at maturity. They suggest that if the IQ is to be regarded as a predictive measure, as has been the custom in the past, the correlation of test items with later rather than with present performance should be taken into account. This is equivalent to saying that in selecting test items the method of sampling should be discarded in favor of one based upon signs. The fact that no such test has yet been standardized may account for the low predictive value of those now available for use with young children.

The significance of an IQ, like any other comparative measure, is dependent upon the frequency of its occurrence. This is ordinarily expressed in terms of the S.D. of a representative sample of subjects. As was stated in an earlier paragraph, the S.D. of a representative sample of white urban school children on the 1916 revision of the Stanford-Binet has been found to be about 16.5 IQ points. Reference to a table of the probability integral shows that this means that in such a sample approximately 68 per cent of the cases will have IQ's that fall between $+1.00$ and -1.00 standard deviations from the mean of the group, or, in this case, between 83.5 and 116.5. Approximately 2.3 per cent will fall beyond the limits of $+2.00$ and -2.00 S.D., with IQ's as high as 133 or as low as 67. Not more than one or two cases in a thousand are likely to

differ from the average by as much as 3.00 S.D., with IQ's as high as 149.5 or as low as 50.5. But if the S.D. of the distribution were greater than 16.5, very high or very low IQ's would be correspondingly more frequent and would therefore be reckoned of less consequence. If, for example, the S.D. were increased by 5 IQ points to 21.5 instead of 16.5, the number of cases with IQ's of 149.5 or higher would be around eleven or twelve in a thousand instead of one or two. Even a relatively small change in the variability of the group means a marked change in the frequency of extreme cases.

It is evident, therefore, that in standardizing a test in which the IQ is to be used as the chief interpretative measure, as much care must be taken to ensure that the variability of the IQ's is the same at all age levels as to see that mental age standards are correct with respect to chronological age. If this is not done, IQ's that are numerically equal cannot have the same meaning at all ages. On the Merrill-Palmer test, for example, it is as easy for a child of 42 months to earn an IQ of 165 as it is for one of 27 months to earn an IQ of 122 (Stutsman, 1931). When the same degree of superiority is expressed by such different figures at varying ages, it is evident, as the author of this test points out, that the intelligence quotient is not an appropriate way of expressing the findings. In spite of this, and in the face of the author's caution, we not infrequently find clinicians reporting IQ's derived from this test, while the calculation of "IQ's" from all sorts of other tests without regard to variability is an unfortunately common practice in many school systems. In spite of its apparent simplicity and the directness of the logic upon which it is based, the intelligence quotient is a very tricky device which should not be tampered with by the statistically illiterate. The uniformity of meaning which is popularly ascribed to it can exist only under certain very rigid conditions which many tests do not meet. Even the 1937 Stanford Revision, upon the standardization of which more time and effort were expended than for any other known test, does not completely meet the requirement of equal IQ variability at all ages (Goodenough, 1942), although McNemar (1942) has published a table of corrective values by the use of which the irregularities may be largely ironed out.

The principle at issue may be stated as follows: If the intelligence quotients derived from a given test are to have uniform meaning at all ages, the standard deviations of the distributions of intelligence quotients must be equal at all ages. It follows as a corollary that this equality can exist only if the standard deviations of the mental ages increase proportionally as age advances.¹⁰

¹⁰ This is evident if it is remembered that the intelligence quotient is merely the ratio between mental age and chronological age, and in order that a ratio may

Another statistical factor of considerable practical importance is the relative stability of high and low IQ's. A change in IQ, when two tests are given to a child at the same age, is entirely the result of a change in his mental age. The same reasoning holds when the tests are given at different ages. Assuming that the conditions specified in the preceding paragraph have been met so that there is no general tendency for the IQ to change with age, the maintenance of a constant ratio between the standard deviations of the mental ages at succeeding ages causes the age factor to cancel out. It follows, then, that the likelihood of change in a child's IQ upon retest is a function of his mental age as well as of his chronological age¹¹ and that among children of the same chronological age, greater variation from test to test may be expected if the initial IQ is high than if it is low. A more complete explanation of this principle has been given by Terman and Merrill (1937), who have shown that in the case of the 1937 Stanford Revision 50 per cent of children with IQ's of 130 or higher may be expected to change their standing by as much as 3.5 IQ points upon retest within a week's interval while the corresponding figure for those with IQ's of 70 or below is only 1.5 points.

Inasmuch as radical changes in the treatment of backward and feeble-minded children¹² are more likely to depend upon the results of mental tests than is the case with children of normal or superior intelligence, the greater dependability of the intelligence quotient at the lower levels is a matter of considerable importance for those actively concerned with the welfare of children. It should be unnecessary to point out, in this connection, that the statistical factors just mentioned may be overridden by poor cooperation, emotional upsets, and other factors which may result in a spurious lowering of the obtained IQ. For this reason it is always necessary to make sure that the test in question was properly given and that the child was in a suitable mental and physical condition at the time of testing. Other things being equal, an IQ is more dependable if it is low, but it must not be forgotten that in an individual case the IQ may have been low because it was undependable.

remain constant, the numerator of the fraction must remain proportional to the denominator. Accordingly if, at age four, the distribution of the mental ages of a representative group of cases has a standard deviation of five mental months, at age eight the corresponding figure should be ten months, and at age twelve it should be fifteen months.

¹¹ The reasons for the low predictive value of tests designed for the early ages have been discussed before. The principle here discussed does not change the relation of test prediction to age but has reference only to the relative stability of high and low IQ's among children of the same age.

¹² Such as placement in special classes or in institutions for the feeble-minded, decisions concerning adoption, etc.

The difference in the reliability of high and low IQ's undoubtedly affords an explanation for some of the disparate reports of IQ change that have appeared in the literature. An examination of these reports will in most cases show that in those investigations where the mean IQ was low the average amount of IQ change was also low, and that the large changes usually have been reported by investigators working with subjects of superior mentality. Other factors, such as the age of the subjects, the interval between testings, and the competence of the examiners, are also to be considered.

SOME REASONS FOR IQ CONSTANCY

If we were to measure the heights of each of a group of 100 young men, all of whom were between the ages of twenty-five and thirty years, and then remeasure them ten years later, the agreement between the two measurements, provided that both had been carefully made, would be very close. The correlation coefficient would certainly be well above $+.90$; in all probability it would be $+.95$ or higher.¹³ The reason is obvious. No further growth in height usually occurs after the age of twenty-five. The second measurement therefore includes all the growth differences made before the first measurement was taken with nothing added or subtracted. The only factor making for difference between the two series of measurements is their inaccuracy.¹⁴ But if the two measurements had been made when the subjects were much younger, say at the ages of two and twelve years, the agreement in their relative standing on the two occasions would have been much lower in spite of the fact that the interval between the two measurements is the same as was used with the men. In this case, however, the height attained at age twelve includes, *in addition to that which had been gained by the age of two*, the very considerable increment in growth made during the interval. Thus only a part of the height measured on the second occasion would have been included in that initially measured; the rest is additional growth.

Suppose that the two measurements had been made at the ages of fourteen and twenty-four years. As in the case just mentioned, the height attained at the later age will include some increment over that which had been reached at fourteen, but the proportion of the final status which is

¹³ Provided, of course, that the range of heights included was as great as is usually found in a sample of that size.

¹⁴ As used here the term "inaccuracy" includes a variety of things, such as differences in posture, in style of shoes (if measurements are taken when subjects are fully clothed), in the kind of measuring instrument, and so on, as well as lack of skill or possible carelessness on the part of the investigator.

accounted for by the growth made between the two measurements is relatively small when compared to the other example.

The relationship between two measures of a growing function thus depends upon the amount of overlap between them as well as upon the experimental error of each measurement. In addition there may or may not be some tendency toward a constant rate of growth for the individual. Children who, at the time of their first measurement, are above the average for their age have obviously grown more rapidly than the generality, either continuously or for one or more periods of time. This initial advantage will be carried over to the time of the later measurement. Even if their growth from then on does not exceed that of the average child, the chances are that they will still be somewhat taller than the generality because of their early precocity.

It is undoubtedly because of this overlap between the amount of growth accomplished previous to the time of the first measurement and that attained by the time of the second that the correlation between earlier and later measurements of any growing function is necessarily positive if the same ability is measured at both times as is presumed in all tests that bear a common name, and if each is measured with reasonable accuracy at the time. It is also assumed that enough growth had been made previous to the first measurement to make it a significant part of the second. If errors of measurement do not exist and the factors measured on the two occasions differ only in a quantitative way, the expected correlation between them is the square root of the proportion that the first is of the second.¹⁵

It is evident, then, that the observed constancy of the IQ does not necessarily connote constancy of individual growth rate as has been affirmed by so many people and as stated as a fact in many textbooks. Such an assumption is not warranted unless it can be shown that the correlation between test and retest exceeds that to be expected on the basis of the overlap between the two after correction is made for unreliability of measurement, or unless other evidence can be adduced in support of the hypothesis. Such evidence might take the form of the correlation between initial standing and gain during the period between the two measurements.¹⁶ If this correlation is positive, indicating that the children who had, on the average, shown the most rapid growth previous to the time of the first measurement also made the most rapid growth

¹⁵ Assuming a rectilinear pattern of growth and equally spaced units of measurement, together with correct location of the zero point.

¹⁶ It is always necessary to correct such a correlation for unreliability of measurement since the uncorrected figure will be far too low unless the errors of measurement are negligible. (Thomson, 1924, 1925; Zieve, 1940.)

during the period between measurements, some evidence in support of the hypothesis that the *rate of growth* tends to be constant for the individual is thereby afforded. But the mere fact that a positive correlation exists between two measurements of the same characteristic when the degree or quantity found at the time of the first measurement is carried over to and becomes a part of that measured later on is insufficient evidence on which to base a theory of constant rate.¹⁷ That the correlation between the IQ's earned by children at earlier and later periods of growth is high enough to justify a prediction of the second on the basis of the first with a degree of accuracy that is considerably better than chance is unquestionably true, and the practical usefulness of such predictions does not hinge upon the more theoretical question as to whether the relationship is entirely the result of the part-whole nature of the data or whether some additional factor making for a constant growth rate for the individual child is also involved. From the scientific point of view, however, the question is an important one which merits much more careful attention than it has received in the past.

The amount of evidence on this head is not great. Anderson (1939) has shown that the correlations actually obtained from several longitudinal studies of the mental growth of children fall short of that to be expected on the basis of the per cent of overlap alone, even after correction for the effect of random errors of measurement has been made. Goodenough (1928) found low positive correlations between initial IQ and gain for 380 preschool children retested by the Kuhlmann (1922) Revision of the Binet tests after an average interval of six weeks, using the Thomson correction formula previously mentioned, but the short-

¹⁷ Anyone who doubts this is invited to try the following very simple experiment. Assume that a group of children are all tested at exactly the same chronological age. Write down a series of mental ages reasonable for that age. Then assume that all are retested after exactly the same interval of time. Take as many separate slips of paper as you have subjects and on each slip write a gain, in terms of mental age, that would be reasonable for the time interval you have assumed. (Ordinarily a child does not lose whatever mental stature he may have already gained any more than he is likely to lose in height, although his gain over a period of time may be negligible or even zero.) Shuffle these slips thoroughly, and add the gains at random to the initial mental ages to get the final mental ages. Then find the correlation between the two series. Since the chronological ages of the subjects were assumed to be identical both at the time of the initial and at the time of the final testing, working in terms of mental ages rather than IQ's will save time in computation and will yield exactly the same results, but those who prefer to reduce the figures to IQ's may do so.

Unless the initial age assumed is so small in proportion to the length of the interval that the ratio of initial to final age becomes very small indeed, it will usually be found that the correlation between initial and final IQ's is about as high as those usually reported in the literature for corresponding ages and intervals between testing, even when, as in this experiment, there has been no tendency to individual constancy of growth rate since gains were assigned at random.

ness of the interval between test and retest makes it uncertain whether these correlations indicate some tendency toward constancy of growth rate or merely greater ability of the brighter children to profit by the first experience.

Another line of evidence pointing to consistency of mental-growth rate has been suggested by Conrad, Freeman, and Jones (1944). As was previously stated, the intelligence quotient can have the same meaning at all ages only if its variability is equal at all ages, and this fact implies that the variability of the mental ages from which it is derived must increase proportionately as age advances. These authors note that the increased variability in mental age necessary for the maintenance of consistent meaning of the intelligence quotient also implies the likelihood that the initially bright are gaining more rapidly than the initially dull and that the dull progress more slowly than the bright. In longitudinal studies where the same subjects are used at all age levels and no factor of selection has been present, this is a reasonable assumption, but when the data are based upon cross-sectional investigations with a different group of subjects at each age level, the use of age variability as a test of consistency of growth rate is more hazardous. However, as the authors of this paper point out, the fact that most of the well-controlled studies on this topic do show some tendency toward an increase in variability with age is in favor of the hypothesis of some degree of rate constancy.

A third line of evidence comes from the study of family resemblances in mental development. All studies that have appeared to date—and there are many—have shown that after the age of very early childhood an appreciable correlation exists between the intelligence test standing of parents and that of their children and an even higher correlation exists between the test scores made by siblings. This could hardly be the case if gains were distributed at random.

Available data, then, appear to favor the assumption that while the chief factor responsible for the observed tendency for intelligence quotients¹⁸ to remain relatively constant is the part-whole relationship, there is some additional tendency toward uniformity of growth rate for the individual. Prediction of future growth rate, however, on the basis of a single IQ is hazardous. The problem of predicting the increment of growth to be made during any specified period of time must not be confused with that of predicting status at the end of that time. The former is based upon the difference between the two measurements and thus is

¹⁸ Or any other measure of relative status within a group in which a part-whole relationship exists.

vitiated by the experimental errors of both, whereas a single measurement is affected only by its own errors of measurement. Of even greater consequence is the fact that the increment must be predicted in its entirety, whereas later status in a growing function is determined to a greater or less extent by earlier status, the amount or level of which has already been determined. It is like betting on the outcome of a race which has already been partly run and the relative position of the participants is known. If the remaining lap is but a small fraction of the total (which is comparable to the situation existing in the greater number of the studies on IQ constancy), the odds against major changes are strong.

Because mental growth as well as physical growth is time-limited, the prediction of later status on the basis of earlier status is facilitated, for the period during which early losses might theoretically be made up is all too short. A child who, by the age of ten has developed mentally only as far as the average child of seven, cannot make up this lack by merely growing at an average rate thereafter. He must increase his pace far beyond that of the average child if his deficiency is to disappear by the time of mental maturity. While this is theoretically possible, experience has shown that it is unlikely to occur. Changes in growth rate do occur, as Bayley (1940) has so admirably shown, and at present there is no really valid evidence to indicate whether such changes occur more frequently at certain ages than at others. It is true that mental status as indicated by the IQ and other interpretative measures shows a marked tendency to "smooth out" as age advances, but as this is a necessary result of the part-whole relationship it affords but meager evidence concerning possible changes in rate of growth.

The intelligence quotient has played an extraordinarily important part in the development of mental testing. To date, no other interpretative measure is so widely used. No other has come so near to being a household word; no other has exerted so strong an influence upon educational practice and sociological thinking. However, the errors arising from its uncritical application to data for which it is unsuited have been many. In this discussion it has seemed well to dwell more strongly upon these errors than upon the unquestionably great service to human welfare that has resulted from its use. The latter is well known and has often been stressed, but in these well-merited eulogies there has been an unfortunate tendency to overlook many of the pitfalls into which the uninformed user may unwittingly stumble. In this chapter, therefore, a few of the more dangerous of these hazards have been pointed out in the hope that those who continue to prefer an interpretation of test

results in terms of the quotient method¹⁹ will be aware of these dangers and so find it easier to avoid them.

But science advances, and in its course new procedures are devised which often tend to supersede the older ones. Valuable as the quotient method has proved itself to be, it is questionable whether it is the best that modern science has developed for the interpretation of test results. In the following chapters, therefore, some alternative methods will be discussed, each of which has some points of advantage over the quotient method as well as special limitations of its own. These include percentile ranks, standard scores, and special derivatives of the latter, such as T-scores and IQ Equivalents, the Heinis Personal Constant, and others.

¹⁹ It may be well to note that intelligence is only one of the characteristics for which the quotient method has been used. Educational quotients based on the results of school achievement tests are very often derived and the same basic procedure is sometimes used for other measures of growth. Thus we find in the literature occasional references to height quotients, social maturity quotients, and a number of others. An article by Rand (1925), although not new, contains many pertinent remarks on the use of the quotient method in educational and psychological practice and these remarks, as well as those in the section just concluded, apply equally to other areas in which this method of expressing test results has been chosen.

Means, Medians, and Percentiles

THE CHOICE OF AN AVERAGE

IN GROUP COMPARISONS

When groups of subjects whose ages cover a fairly wide range are studied, the possible irregularities in the quantitative meaning of interpretative scores introduce a number of problems. Suppose, for example, that a test standardized by the year-scale method, such as the Stanford-Binet, has been used, or that mental ages and intelligence quotients have been worked out on the basis of some group test. The five cases below illustrate one of the difficulties:

Chronological age (in months)	Mental age (in months)	IQ
50	75	150
80	60	75
40	50	125
90	60	67
30	25	83

The mean chronological age is 58 months, the mean mental age is 54 months, but the mean IQ is 100. The reason for this apparent discrepancy lies in the fact that the mean of a series of ratios will be the same as the ratio of the means of the terms upon which the ratios are based *only if the bases of all the ratios are equal*—in this case, if all the chronological ages had been the same. It thus becomes necessary to make certain assumptions whenever data of this kind are to be handled, and it is important for all who make such calculations, however simple and direct they may seem, to know what these assumptions are.

If one finds the mean of a series of IQ's when the subjects differ in age, he makes the implicit assumption that IQ's have the same meaning and are equally significant at all ages. He assumes further that for him

the important thing to ascertain is the relative degree of brightness of his subjects and not the level of mental maturity they have attained. The same assumption holds, of course, for educational quotients or any other maturity quotients which he may obtain. The question of what is the most significant score must be answered by the individual investigator. It depends upon the particular problem which he has set out to solve. But the question of the legitimacy of summing and averaging IQ's is a different matter and one that hinges upon the points discussed in the last chapter. If, as in the case of the Merrill-Palmer tests, the requirement of equal standard deviations of IQ at all ages is not fulfilled and numerically equal scores therefore have very different quantitative meaning, the averaging of scores is obviously not permissible from any standpoint, either scientific or practical.

What about finding the mean of the scores themselves, before transmuting them into IQ's? If scores are expressed in truly equal units, this is quite as permissible as finding the average height or weight of a group, and has similar limitations of meaning. For if children differ greatly in age, there are relatively few problems for which the determination of their average height is of much significance. Moreover, since growth in height is not a straight-line function except for relatively short periods of time (see Figure 7, page 159), the ratio of average age to average height also involves a mathematical error since the terms of the ratio do not progress in a parallel manner. In mental tests standardized by the year-scale method, the inequalities of mental growth from age to age which probably exist have been artificially ironed out by placing an equal number of tests at each age,¹ but this does not completely solve the difficulty. For the difference between the mental and the chronological ages of a young child whose average rate of mental growth is indicated, let us say, by an IQ of 133 will be much smaller than that of an older child who has maintained the same average rate for a longer period of time. Accordingly the standard deviations of the distribution of mental ages will be less affected by the inclusion of some young children with very high or very low IQ's than will be the corresponding standard deviations of the distribution of IQ's, since a child of twelve years must have attained a mental age four years in advance of his chronological age in order to attain an IQ of 133, whereas the child of three years needs only a single year's excess in order to reach a similar IQ level.

Since the IQ's of older children are more dependable in the sense

¹ This is equivalent to dividing the absolute scale of mental growth into finer units during those periods when growth is slow and the absolute amount of gain is correspondingly small, and using coarser divisions during periods of rapid mental growth.

that they predict mental status at maturity more accurately than do those obtained from young children, there may be some justification for using the ratio of the average mental age to the average chronological age as an indication of the most typical mental level of a group whose chronological ages vary, when the test used has been standardized by the year-scale method. It should be clearly understood, however, that when this is done more weight is assigned to the mental standing of the older children in the group than to that of the younger ones. If the opposite method, that of averaging the individual IQ's, is employed, the assumption that the IQ has similar meaning at all ages and that individual IQ's remain relatively constant over the range of ages covered in the study must be warranted.

When, in the light of these facts, neither the mean IQ nor the ratio of the means of the terms from which the IQ is derived seems appropriate to the needs of a given problem, the only remaining alternatives are (1) to utilize some other interpretative measure for expressing the results of the tests, or (2) to limit comparisons to children of the same or closely similar chronological ages. If the data actually include a range of ages, the results may be subdivided on the basis of age, even though the number of cases in each age comparison is necessarily reduced by such a division.

As a rule, the arithmetic mean obtained by dividing the sum of the scores by the number of cases in the group² is the most dependable measure of central tendency, that is, the measure which varies least from sample to sample of a given universe. But in some instances when the data chance to include a number of cases with scores that diverge so much from the generality that their inclusion raises considerable doubt as to the representativeness of the sample in which they occur,³ the median,

² The statistical short cut known as the *method of the guessed mean*, which is described in any of the elementary textbooks on statistical methods, is an alternative often employed when the number of cases exceeds fifteen or twenty.

³ The representativeness of a sample is judged on the basis of the agreement between its mean and standard deviation and the corresponding measures of the universe from which it is drawn when the latter facts are known. In many cases, however, these functions have not been or cannot be ascertained for the total universe, since the latter may be of infinite size or at any rate too large for feasible handling. In such cases the representativeness of a given sample may be determined either empirically by comparison with the results of a number of other samples drawn in a similar manner, or statistically in terms of the size of the sample and the amount of variation within it. From these two facts (or statistics) the most probable magnitude of the standard deviation of the distribution of the means of an infinite number of samples drawn from the universe in question—in other words, a measure of the amount of confidence which may be placed in the mean of a single sample—can be obtained. Such a standard deviation, which is based upon theoretical rather than wholly empirical data, is known as the *standard error* of the measure to which it

rather than the mean, may give a better picture of the facts. The median is the point above and below which exactly 50 per cent of the cases will fall. In other words, it is the point which divides the subjects into two numerically equal groups after they have been arranged in rank order according to their scores on some test or measurement. In making this division, no account is taken of the *distance* between the point of division (the median position) and the position of any individual. One whose score is fifty points removed counts for no more than one who differs from the median by only a single point, provided that the difference is in the same direction. It follows that the median is less affected by the inclusion of a small number of extreme cases than is the mean. It also follows that changes occurring near the midpoint of the distribution are likely to have a greater effect upon the median than on the mean because of the likelihood that such changes will mean a shifting of some persons originally falling in the lower half to the upper, and vice versa. In a normal distribution, therefore, the sampling error of the median is ordinarily greater than that of the mean⁴ and in a platykurtic distribution (an unusually flat curve in which the tendency for the greater number of the cases to cluster about the midpoint is less marked than is generally the case) this is even more true. But if the reverse condition exists, with

refers (in this case, as the *standard error of the mean*). It is thus a measure of the *theoretical variability of the means of a series of different samples drawn from the same universe*, just as the ordinary standard deviation is a measure of the amount of variability within a single group or sample about its own mean. The latter is an empirical measure, however, based upon the facts obtained for an actual distribution of real cases; the former is merely a matter of statistical probability in which, although all the known data are taken into account, the greater part of the data remains unknown. The standard error of a statistic is the best estimate of its dependability that can be had from the data at hand, but it is an approximation based upon a measurement rather than a measurement in the strict sense of the word.

The current practice of employing the same written expression (the lower-case Greek σ) for the standard deviation of an actual distribution of numbers and for the theoretical distribution of the statistics of a series of samples drawn from the same universe seems to me to be deprecated because of the unnecessary confusion thereby produced in the minds of inexperienced persons. We shall therefore use the term "standard deviation" (S.D.) only in referring to the variability within a given sample, and reserve the term "standard error" (σ) to denote the theoretical variability of a series of samples from the same population, and hence the probable stability of a given measure of the sample in question.

⁴ If the distribution is exactly normal, the standard error of the median will be approximately 1.25 times that of the mean. Some elementary textbooks on statistical methods make the statement that this is a universal relationship. Their rule for determining the standard error of the median is first to find that of the mean and then multiply the result by 1.25. The reader should understand, however, that this method is legitimate only in those cases in which the distribution of scores is normal or nearly so, and that there are instances in which the median is preferred to the mean for the very reason that it is more stable, i.e., has a smaller standard error. For a more general method of finding the standard error of the median see Kellev, *Statistical method* (1923).

most of the measures centering around the midpoint fairly closely but with a few trailing out rather far toward the extremes, so that the impression given is that of a very high central peak with long tails at either side of it, the median may actually be the more stable of the two.⁵

Sometimes, however, one is interested in knowing at what point in a distribution most of the cases congregate. When only the cases in a particular sample are to be considered, this point is known as the *crude mode*; if a statistical estimate is made of the most probable point of greatest concentration in the universe from which the sample is drawn, this position is known as the *true mode*. Unless the distribution is very nonsymmetrical, with most of the cases congregated at one end but the remainder stretching out in a long tail to form what is known as a *skewed curve*, a good approximation to the true mode may be had by applying the formula given below:

$$\text{Mode} = M - 3.03 (M - Mdn),$$

where M is the mean and Mdn the median of the sample distribution (Kelley, 1923).

The mode is not much used in research because it is less stable than either the mean or the median, and does not enter into other statistical computations likely to be made. The crude mode is sometimes important for the teacher or the social worker whose interest centers about a particular sample of cases and who has relatively little concern with the universe of which it is a part. This is particularly likely to be the case when the scale of measurement employed is divided into only a small number of coarse steps and a large proportion of the cases are given the same ranking.

One other measure of central tendency should be mentioned: the *harmonic mean*, which is defined as the reciprocal of the mean of the reciprocals of the separate measures. Its chief application to the field of mental testing is in those cases where it is desired to secure a measure of rate from data which have been scored in terms of amount accomplished in a given period of time. For example, a certain test of motor speed for young children involves the carrying of small blocks from one side of the room to another position six feet distant. The most convenient way of scoring this test is in terms of the number of blocks moved in a given short period of time, and one way of expressing the group average is to take the mean number of blocks moved during this specified period. Sup-

⁵ The statistical term which refers to the relative frequency of scores at the different points of a distribution is known as *kurtosis*. If flatter than usual, a distribution is said to be *platykurtic*; if or about average flatness (normal), it is called *mesokurtic*; if unusually peaked at the center with long tails caused by the occurrence of a few atypical scores, it is said to be *leptokurtic*.

pose, however, that one were interested in studying the work decrement (due to fatigue, loss of interest, and so on) over a period of time. In this case the average change in rate of work is probably more significant than the average change in amount accomplished, and accordingly the harmonic mean of the successive work periods is probably a more significant indication of the extent of the work decrement than would be the arithmetic mean of the corresponding changes in amount accomplished per unit of time.

In choosing one of the various ways of securing an indication of the most typical performance of a group of subjects, that is, a measure of the central tendency of the group, the first and by far the most important consideration is the exact nature of the question to be answered. The mean takes account of all the measures, giving proportional weight to extreme cases. The median is simply the midpoint from which large divergences count for no more than small ones taking the same direction. The crude mode shows the point of greatest concentration in the sample; by statistical computation an estimate can be had of its most probable position in the universe from which the sample is drawn. The harmonic mean provides a method of transforming data expressed in terms of amount accomplished per unit of time into units of rate, i.e., time required per unit of work. No one of these measures is universally to be preferred, although the fact that in the majority of cases the arithmetic mean is the most stable and is used in so many other statistical formulas lends it some advantage. But the crux of the matter is the nature of the problem to be answered. Other points, such as the relative stability of the measures as indicated by the magnitude of their standard errors, may be taken into account when either of two methods seems equally suitable.

PERCENTILE RANKS AS INTERPRETATIVE MEASURES

A method of expressing the relative position of an individual within the group to which he belongs that has attained much popularity within recent years is through the use of percentile ranks. Just as the median marks the point above or below which 50 per cent of the cases fall, so the point attained by any other stated percentage of a given group may be determined and used as a means of interpreting the scores made by individual members of the group. In utilizing this method it is necessary first to establish percentile points⁶ marking off the levels attained by

⁶ If sufficiently large groups are available (say a thousand cases or more), this may be done empirically for each separate percentage. If the amount of data is too small to permit so fine a grouping, coarser units such as successive 5 per cents or

each successive percentage of the group in question. It then becomes possible, by comparing the score made by an individual with the table of percentile values, to interpret this result in terms of the percentage of his group whom he equals or surpasses. For example, if John Doe earns

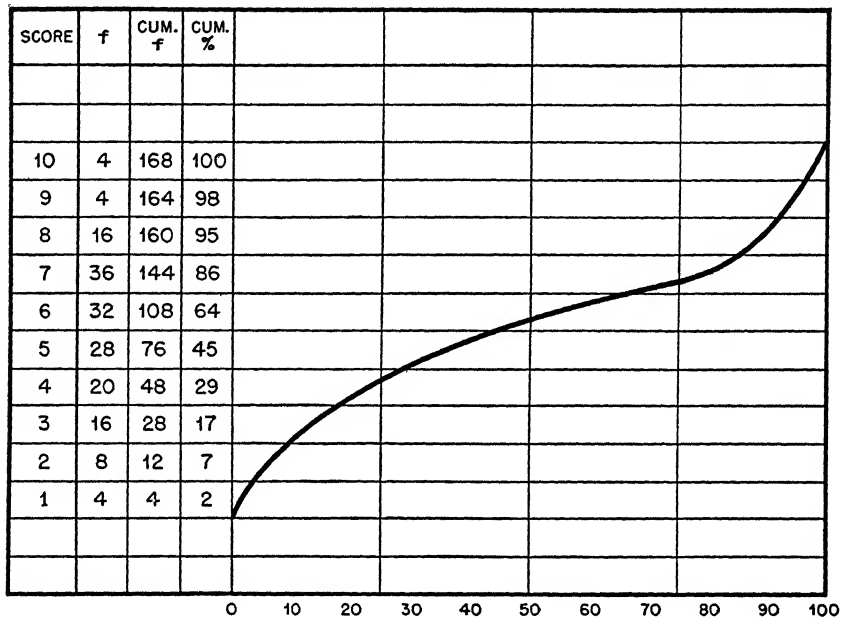


FIG. 8. A PERCENTILE CURVE.

a score of 16 on a given test, and if this score is exceeded by 98 per cent of the group with whom he is classified—in other words if there are only 2 per cent whom he equals or surpasses—he is said to have a percentile rank of 2. But if, on another test with different standards, he also earns a score of 16 but in this case 75 per cent of the group do no better than he, his percentile rank on the second test would be 75.

The percentile method thus provides another way by which scores on a variety of different measures scored in different units and with different numerical standards may be reduced to uniform and meaning-

even 10 per cents (deciles) may be used to mark off the successive points, although this necessarily involves some loss in accuracy. By interpolation between the points so indicated, taking account of the fact that a percentile curve is not a straight line but usually involves a double flexure (see Figure 8), a graphic indication of the successive percentile values may be had from which a table for use in interpreting individual scores may be drawn up.

ful terms.⁷ Its meaning is readily grasped even by persons without training in statistical methods or advanced mathematics. As compared to the intelligence quotient it has the great advantage of making fewer assumptions, and in its best tradition—which unfortunately is sometimes disregarded—of stating clearly just what these assumptions are. For the use of percentiles does not necessarily imply that a representative sample of a major universe has been obtained as is the case with mental ages and intelligence quotients. A percentile rank usually is, and always should be, specified as referring only to a given sample, such as the ten-year-olds enrolled in the public schools of a given city during a certain school year. Moreover, in the reporting of normative standards for use in individual classification, a sound custom is being established of stating whether the entire universe was examined as a basis for these standards or, if a presumably representative sample was used, just what precautions were taken in order to avoid bias in the selection. Such information greatly lessens the likelihood of errors arising from the comparison of an individual with standards derived from a group to which he does not belong. When immature individuals are involved, the fact that comparisons are made only between subjects of the same chronological age⁸ does away

⁷ There is no really logical principle to determine whether the counting should begin at the upper or the lower end of a given distribution; i.e., whether a percentile rank of 5 should mean that the subject in question exceeds or is exceeded by 5 per cent of the group to which he belongs. When the percentile method first came into use, a good deal of confusion arose from the fact that some investigators used the first method of computing percentiles, others the second, which made it necessary always to ascertain the method of computation before one could be sure whether, for example, an individual who had been assigned a percentile rank of 2 on an intelligence test was exceeded in brightness by only 2 per cent of his age group or whether only 2 per cent were equally stupid. Although modern usage is pretty well agreed on the second of the two possible directions in making the count, some confusion still exists, especially with respect to such tests as the so-called "personality inventories" and others in which the usual direction of values is reversed, making a high score undesirable, a low score desirable. Research workers actually using the tests are unlikely to go astray on these points, but teachers, social workers, and others who attempt to make use of the scores reported for their clientele and who are not themselves well versed in statistical methods may be seriously misled by such inconsistencies. Another source of frequent confusion for the practical worker is the fact that the numerical values of percentile ranks do not correspond to those of the IQ with which many of them are more familiar. A percentile rank of 50 corresponds to the median and thus indicates average standing. An IQ of 50, on the other hand, if it is valid, usually means mental deficiency of a level often requiring institutionalization. I have personally known of a fair number of instances in which percentile ranks have been entered on the permanent record cards of school children as "IQ's," with consequent gross underestimation of their level of ability, and of at least one case where a normal child was judged to be unsuitable for adoption as a result of the same kind of clerical blunder.

⁸ If the function measured is changing rapidly with advancing age, it is important to keep the age intervals short enough so that the younger children in a normative group will not be penalized by their mere immaturity nor will the older ones be given an unfair advantage by reason of their greater age.

with a number of the problems arising from the unsettled question as to the form of mental-growth curves when plotted against time. Percentile ranks also have the advantage of being as suitable for adults as for children. Properly applied, they permit a kind of group comparison which is fundamentally more sound than many of those made on the basis of mental ages or intelligence quotients, but it should be noted that comparisons of this kind which are mathematically indefensible are unfortunately all too common. For *percentile ranks should never be averaged*, either for an individual or for the members of a group. If groups are to be compared, the correct procedure is to find percentile standards for each comparison group separately and then compare these standards. If, for example, it should turn out that 50 per cent of Group A achieves a level which is reached by only 5 per cent of Group B, that fact is meaningful since the statistical probability of its maintaining the same direction of difference in repeated samples from the same universes can be determined. But it is not permissible to find the percentile ranks of the subjects in each of the two groups by comparing their scores with a standard percentile table and then averaging the ranks in each group for purposes of comparison, as is sometimes done by those unacquainted with the particular limitations of the percentile method.

The most important of these limitations can readily be seen by comparing Figures 9 and 10. Since percentiles are determined by the proportionate number of cases who earn scores falling within a stated range, a given percentile is represented on a distribution curve by a certain area of that curve. And it will require only a glance at Figure 9 to demonstrate that if percentages of the total area of this curve or of any other that roughly resembles it are to be kept equal, the distances subtended by these equal areas along the base line of the curve cannot be equal. Only in the very unlikely case represented by Figure 10, in which the form of the distribution is rectangular, with as many cases at the extremes as at the midpoint and with the measures beginning and ending abruptly at certain sharply established boundary lines, is it possible for equal percentiles to correspond to equally spaced units of measurement. Since this is true, it is apparent that percentile ranks cannot justifiably be handled by the ordinary arithmetical processes. They should not be added, subtracted, multiplied, or averaged. As interpretative measures they have much value, both because they are so readily understood and because, when properly used, they are free from some of the overgeneralizations likely to be inherent in age and quotient methods, where the implication of some kind of universal reference is strong but the limits of this universe are rarely specified, and where the growth curve is assumed to correspond to the time curve either absolutely or in terms of proportional

variability. The use of percentile ranks also obviates some of the difficulties arising from the correlation between developmental factors when chronological age is not controlled. A mental age, an educational age, or a social maturity age of ten years has not the same meaning for a

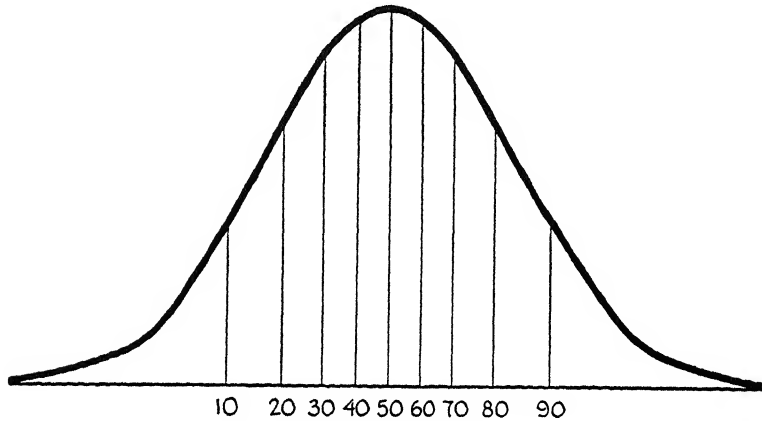


FIG. 9. UNEQUAL DISTANCES BETWEEN POINTS SUBTENDED ON THE BASE LINE OF A NORMAL CURVE BY SUCCESSIVE 10 PER CENT DIVISIONS (DECILES) OF ITS AREA.

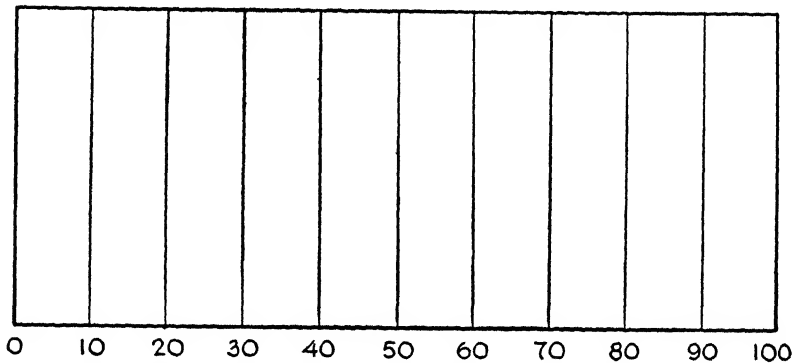


FIG. 10. EQUAL DISTANCES BETWEEN POINTS SUBTENDED ON THE BASE LINE OF A RECTANGLE BY SUCCESSIVE 10 PER CENT DIVISIONS (DECILES) OF ITS AREA.

child of six as for another of sixteen. Too many other factors which change with chronological age alter the picture. To a lesser extent the same rule perhaps holds for children from very different social backgrounds. Moreover, the practical significance of comparisons between individuals of noncompeting groups is likely to be less than that of comparisons between individuals belonging to the same group. Generally speaking, one is more interested in knowing how an individual stands

among the persons with whom he associates than in ascertaining his status within a universe so large that he will never come into contact with most of its members.

THE STABILITY OF PERCENTILES

Reference to Figure 9 will show that as one moves out from the center of a symmetrical curve of distribution toward the extremes, the height of the ordinate which represents the number of cases making a given score steadily decreases. In other words, the assignment of a percentile rank to a given score is based upon smaller and smaller samplings of cases as scores become more remote from the midpoint.⁹ This is equivalent to saying that, in general, percentile standards are more reliably established at the median than at the extremes. When ascertaining the most probable extent of variability of a percentile, the following formula given by Kelley (1923) for determining the standard error of a *percentile* is always to be used rather than that for determining the standard error of a *percentage*, which is sometimes erroneously substituted for it:¹⁰

$$\sigma_{\%ile} = \frac{i_{\%ile} \sqrt{Npq}}{f_{\%ile}}$$

where $i_{\%ile}$ is the interval covered by the percentile in question expressed in score points;

N is the total number of cases in the group upon which the standards are based;

p and q are the percentages lying above and below the point of dichotomy (together making up 100 per cent); and

$f_{\%ile}$ is the number of cases in a percentile class ($1/100 N$ if single percentiles have been used in making the computations).

OTHER MEASURES BASED ON PERCENTAGES

When only relatively coarse classification is called for, the points marking the scores made by successive tenths of the normative group may serve the purpose. These positions are known as *decile points*, and a subject whose score falls within the ranges so defined is said to score within that decile. If the scores of the lowest 10 per cent range from 0 to 14 on a

⁹ This statement is true for all unimodal curves that are not definitely skewed. In skewed curves the change in stability will be more rapid toward the long than toward the short tail, and in multimodal curves special formulas are needed to cover each case.

¹⁰ The standard error of a *percentage* is given by the formula

$$\sigma_p = \sqrt{\frac{pq}{N}}.$$

particular test, while those of the next 10 per cent range from 15 to 22, a subject whose score is 18 would be classed in the second decile. If only a very crude classification is required, the *quartiles*, which are the points marking off the scores made by the lowest 25 per cent, the median (the 50 per cent point), and the best 25 per cent, are sometimes used.

If, on the contrary, an exceedingly fine system of classification is wanted, *permilles* are sometimes substituted for percentiles. Permilles, as the name implies, indicate the points marked off by the performances of successive thousandths of the group in question, just as percentiles indicate the positions of successive hundredths of it.

All of the above measures are, of course, merely special points on a percentile distribution and are subject to the same limitations as other percentile measures. All are unequally spaced along the base line and consequently are not subject to ordinary arithmetical treatment. Like those of other percentile measures, their meaning is easily understood by persons with little technical training, while the common practice of specifying the universe to which they apply¹¹ has the great advantage of avoiding overgeneralization by setting boundaries within an area where such terms as "the general population" or "the average six-year-old" have too often roamed unmolested and unchallenged.

¹¹ As, for example, "He ranks within the top quartile of the senior class at Blank High School," or "Her score on the X test fell within the second decile of entering freshmen at this college."

Standard Scores and Their Derivatives

In the last chapter we noted that the great advantage of the percentile rank as an interpretative measure lies in the ease with which it may be comprehended by persons who wish to make practical use of test results but who have had little or no training in the more theoretical aspects of test construction. We turn now to another method of reducing scores to a uniform basis which is perhaps more difficult to grasp at the outset but, once comprehended, is more easily handled. Unlike percentiles, standard scores mark off equal distances along the base line of a distribution curve¹ and therefore are subject to arithmetical treatment. They may be added, subtracted, divided, or averaged. Finally, as will be shown in a later section of this chapter, they may readily be converted into other and more familiar terms for use in explaining test results to nurses, teachers, and other practical workers with children.

As immediately derived, a standard score is the number of standard deviations by which the unconverted or "raw" score on a given measure exceeds or falls below the mean score earned by the group to which the subject who makes the score belongs. Inasmuch as deviations are calculated in both directions from the mean as zero, it is necessary to prefix a plus or a minus sign to indicate whether the score in question has a positive or a negative value. A standard score of $+1.5$ would thus be $1\frac{1}{2}$ standard deviations above the average of the group. If the scores for this group are normally distributed, this corresponds to a percentile rank of approximately 93. Conversely, a standard score of -1.5 is $1\frac{1}{2}$ standard deviations below the mean of the group. In a normal distribution the corresponding percentile rank is close to 7.²

Because it is not always convenient to deal with both negative and

¹ Assuming, of course, that the units along the base line are equally spaced. In the practical situation, minor irregularities in spacing of units will to some extent be smoothed out by the standard-score method of treatment, provided that these irregularities are distributed more or less at random and do not tend to cluster at certain points.

² By consulting a table of the normal probability integral (see p. 195), percentile ranks may be read directly from standard scores and vice versa, provided that the distribution is normal or approximately normal.

positive numbers and also because their use often leads to computational errors, several ways of avoiding the negative sign have been proposed. One way is to change the position of the zero point from the mean to a point sufficiently far below the mean to include all except exceedingly rare cases.³ Inasmuch as in a normal distribution only about one case in a million will fall below the distance marked off by a point 5 standard deviations below the mean, the addition of a constant value of 5 S.D. to all the scores would not change their relation to each other but would bring practically all above the zero point, thus doing away with negative values. By this system, a score that is one standard deviation below the mean becomes 4 ($-1 + 5 = 4$) and one that is one standard deviation above the mean becomes 6 in the new notation. Since all signs have now become positive, they need no longer be recorded.

If the distribution of original scores is definitely skewed, however, and there is reason to think that a normal distribution gives a better picture of the facts,⁴ a correction must be made to the raw-score values before proceeding to the computation of standard scores.

³ The tails of a theoretical distribution of a universe with indeterminate boundaries stretch out to infinity. It thus becomes impossible to locate a point above which all conceivable cases would fall.

⁴ There are two possible conditions in which this may be the case. In the first, the obtained distribution of scores may be regarded as correct from one point of view but not from that in which the experimenter is interested. For example, an investigator may wish to use some measure of economic status in connection, let us say, with a study of factors correlated with individual differences in child intelligence. Family income seems to be the best figure for his purpose but, as is well known, the distribution of incomes is very greatly skewed, with the vast majority of cases massed toward the lower end and a small number tailing out to a point so far above the median level that they cannot be shown on a graph of ordinary dimensions plotted in the usual way. For the economist these extreme incomes are meaningful, but for the educator or the psychologist who is concerned with the welfare of children, not much practical importance attaches to the difference between an annual income of \$25,000 and one of \$250,000 or more. Certainly the difference here, when conceived in terms of what worth-while things can be done with the money, is far smaller than that resulting from the difference between an income of \$1000 a year and one of \$2500. Any method of scaling which is based upon absolute terms of dollars and cents will obviously give a very incorrect picture of the facts considered from the standpoint of the kind of living conditions made possible by a given level of income.

In the second instance, there may be reason to believe that the skewing is wholly an artifact, resulting either from bias in the instrument itself or in the manner of using it, or from its application to a group for whom it is either too easy or too difficult. The latter condition was discussed in Chapter 10. Bias in the instrument may occur because the test maker was inept in devising a sufficient number of items at either the easy or the difficult end of his scale. As a result, the distances from one item to the next will be greater at one end of the scale than at the other. Of course if a suitable method of scaling the individual items has been used, such discrepancies should not occur since the irregularities in calibration would at once become apparent. But if a simple item-count method is used, and the spacing of items tends to become farther and farther apart as one moves from one end of the scale to the other, that is, if the scaling errors show a constant trend instead of being distributed at random, a calculation of

Consider the question of the relation of income to standard of living for the individual family at any given period of time. There are, of course, certain differentials for which at least a rough correction must be made—size of family, place of residence, and so on. Other differences, such as thrift in managing the family budget, also exist, but as these factors are rarely known with sufficient accuracy to permit making allowances for them, it will usually be necessary to regard them as random errors which will tend to cancel each other if the group studied is sufficiently large. Now if the interest in a particular investigation centers about the question of the material advantages that a given level of income makes possible, and if a series of figures showing the relative status of each individual on such a continuum is needed, a calculation of the mean and standard deviation of actual income figures for a representative sample of the population would yield spuriously high figures, and standard scores derived from these results would be equally misleading. Percentile ranks might be used, since no assumption is thereby made as to the form of the distribution. The disadvantage of this method, as was pointed out before, lies in the fact that percentiles are not subject to further mathematical treatment. But if the underlying assumptions are understood, it is possible to compute standard scores for a long-tailed distribution such as is given by the income figures, and to base these scores upon as fine or as coarse a grouping of the data as seems desirable. What is evidently needed is the standard deviation value (standard score) of a level of income attained by a given percentage of the population. The procedure, which has been outlined in detail by Kelley (1923) and by Peters and Van Voorhis (1940), is not difficult to understand in principle though its derivation is somewhat more complicated. Let us consider the *Report of the Dominion Bureau of Statistics* on Canadian incomes during the year 1942 as given in Table 3.

As they stand, the data cannot readily be plotted in the form of an

the standard deviation which is based upon the uncorrected scores will be in error by an appreciable amount, and the standard scores derived from it will be even more seriously biased, since the true deviations of the individual scores from the true mean will be underestimated at one end and overestimated at the other. Bias in the manner of using a scale may also occur if the method of obtaining the scores is not wholly objective. For example, the letter-grade system of marking employed in many schools and colleges represents a scale of merit in which descriptive meanings are assigned to each letter used. Some teachers, however, are reluctant to give low grades. Their distributions of marks will usually show an excess of A's and B's with a corresponding deficiency of marks below C. A grade of A in such a case appears to mean more than it actually does if one considers only the official description; a grade of D or below really indicates greater deficiency than appears. If the marks of such a teacher are to be brought into line with those of others who follow the official descriptions more literally, a normalizing process must be applied to the grades before such a combination is warranted.

TABLE 3

DISTRIBUTION OF CANADIAN INCOMES FOR THE YEAR 1942

<i>Amount of income</i>	<i>Per cent of cases</i>
Under \$500	18.4
\$500- 1,000	26.5
\$1,000- 2,000	39.0
\$2,000- 5,000	14.5
\$5,000-10,000	1.1
\$10,000-25,000	0.3
Over \$25,000	0.1

ordinary line graph for two reasons. Inasmuch as more than 80 per cent of the incomes do not exceed \$2000, a reasonably fine scale must be used for plotting if distinctions within the main body of the data are to be

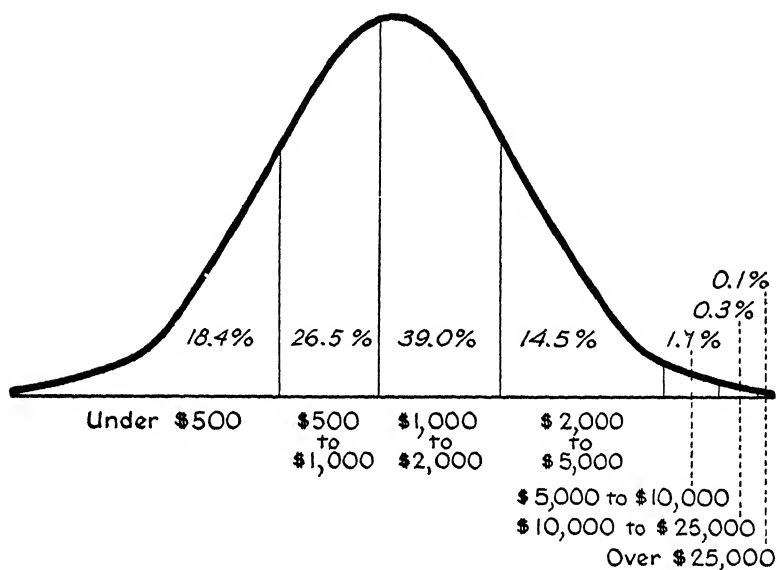


FIG. 11. SHOWING THE DATA OF TABLE 3 AFTER THE VALUES HAVE BEEN TRANSMUTED INTO THE FORM OF A NORMAL DISTRIBUTION.

shown. But the use of such a scale would require an inconveniently long space in order to include the upper levels at all. Moreover, the facts as given are not sufficiently detailed to permit plotting the curve within the separate income classes with any degree of accuracy. The decrease is evidently very sharp within the range from \$2000 to \$5000, but it is unlikely that a straight line would express the facts.

Figure 11 shows the data of Table 3 in the form of a normal distri-

bution. It will be noted that no actual violence has been done to the facts as presented. The difference consists in regarding the frequency of the various income classes, rather than their absolute size, as the essential facts. It may be argued that this is quite as legitimate a way of regarding the matter as the other would be, since in any society income is largely a comparative matter.

For the details of the method the reader should consult the references given. Certain points should, however, be noted. First, although the method assumes that the facts reported represent a numerical continuum in which the successive classes are arranged in serial order, it is unnecessary that these classes cover equal ranges of the continuum, or that the number of cases in each class be equal. All that is needed is a knowledge of the percentage of the total included in each class. These percentages are represented by the *areas* of the successive segments. Just as a cake may be cut into large or small slices at will, so these segments may be large or small according to the percentages of the total which they represent. The division points are read directly from a table of the probability integral which may be found in any standard textbook on statistical methods. A much condensed form of this table is given on page 195.

The problem now becomes that of finding the most representative single value to be assigned to each income class. It is apparent at once that the mid-value of the class would not be the correct figure to use, since so many more of the individual incomes would fall on the side nearer the midpoint of the group than on the side farther away from it. Calculation from the probability integral gives the necessary figure in standard scores.⁵ It would be possible, if desired, to translate these values back into their original units, but as a rule little would be gained from doing this.

It may be well to repeat that the legitimacy of transmuting data into standard score units (or any other derived values) is a matter of the assumptions one is willing to make with reference to the problem at hand. Whether one is aware of it or not, every procedure that is carried

⁵ The formula is

$$d = \frac{z_1 - z_2}{q_1 - q_2} \quad (\text{Kelley, 1923, pp. 99-101})$$

where d is the standard score required;

z_1 and z_2 are the ordinates of the normal curve at the upper and lower boundaries of the segment in question (read from the table);

q_1 is the proportion of the total area of the curve falling beyond its *upper* boundary; and

q_2 is the proportion of the total area falling beyond the *lower* boundary of the segment (q_1 + the proportion included in the segment itself).

out involves its own set of underlying assumptions, and this statement applies as much to "leaving the data as they stand" as to transforming them into other units. One may quite as properly ask whether an experimenter is justified in omitting to make a statistical transmutation of original units into some kind of derived scale as whether or not it is permissible to do so. No hard and fast rule can be given; the answer depends on the nature of the original data and on the problem. The essential thing is to realize that the question always exists and cannot safely be ignored.

Many different applications of the standard score method occur in the field of mental testing, which in general will be discussed in connection with the problems to which they apply. Brief mention may be made here of some of them. The difficulty value of an item for a particular group may be found very simply from the probability table by locating the position on the base line (marked x) of the point dividing the percentage of those who are able to succeed with the item from the percentage failing it. If fewer than half succeed, the standard score will have a plus sign, indicating that it is above average in difficulty. If more than half succeed, the score will have a minus sign. Suppose, for example, that only 20 per cent succeed. The point marking off the upper 20 per cent of the area is located 0.84 S.D. above the mean. The standard score value of that item would therefore be $+ .84$.

The standard score method, with some minor variations, is also used extensively in the construction of scales of merit based upon the judgments of large groups of raters and in the allied problems of constructing attitude scales, interest scales, and other measuring devices which do not show a developmental trend or in which the maturity value is secondary to the main purpose of the scale.

Except for the fact that they are slightly more difficult to explain in nonmathematical terms, standard scores share in most of the advantages that have been noted for percentiles and have few of their disadvantages. That they are subject to mathematical treatment has already been mentioned. Since the calculation of the S.D. is based upon all the cases with proportional weighting of the extremes, standard scores are equally stable at all points in the distribution. This, it will be recalled, is not true of percentile ranks. Like percentiles, standard scores are generally computed only for groups that are homogeneous with respect to age or other variables likely to introduce unknown and perhaps spurious⁶ factors into the results. In a few cases test makers using the standard

⁶ The term "spurious" as used here has reference only to the factor which it is desired to measure. A factor which is spurious in one connection may be valid and important in another.

TABLE 4
TABLE OF THE NORMAL PROBABILITY INTEGRAL AT SUCCESSIVE
1 PER CENT LEVELS*

<i>q</i>	<i>x</i>	<i>z</i>	<i>q</i>	<i>x</i>	<i>z</i>
.50	.000	.399	.20	.842	.280
.49	.025	.399	.19	.878	.271
.48	.050	.398	.18	.915	.262
.47	.075	.398	.17	.954	.253
.46	.100	.397	.16	.994	.243
.45	.126	.396	.15	1.036	.233
.44	.151	.394	.14	1.080	.223
.43	.176	.393	.13	1.126	.212
.42	.202	.391	.12	1.175	.200
.41	.228	.389	.11	1.227	.188
.40	.253	.386	.10	1.282	.175
.39	.279	.384	.09	1.341	.162
.38	.305	.381	.08	1.405	.149
.37	.332	.378	.07	1.476	.134
.36	.358	.374	.06	1.555	.119
.35	.385	.370	.05	1.645	.103
.34	.412	.366	.04	1.751	.086
.33	.440	.362	.03	1.881	.068
.32	.468	.358	.02	2.054	.048
.31	.496	.353	.01	2.326	.027
.30	.524	.348	.009	2.366	.024
.29	.553	.342	.008	2.409	.022
.28	.583	.337	.007	2.457	.019
.27	.613	.331	.006	2.512	.017
.26	.643	.324	.005	2.576	.014
.25	.674	.318	.004	2.652	.012
.24	.706	.311	.003	2.748	.009
.23	.739	.304	.002	2.878	.006
.22	.772	.296	.001	3.090	.003
.21	.806	.288			

* This table gives values for one side of the distribution only. The column headed *q* represents the smaller of the two proportions which together make up 100 per cent of the distribution. The second proportion (*p*) may be found by subtracting the value of *q* from 1.00; thus where *q* = .25, *p* = .75, etc. Since the table is symmetrical, the values of *x* and *z* will be the same for *p* as for corresponding values of *q*.

The column headed *x* represents the number of standard deviations above or below the mean corresponding to the proportion indicated by *q*. Its sign is either + or - accordingly as the point of dichotomy falls above or below the mean.

The height of the ordinate at the point of dichotomy is shown in the column headed *z*.

The table serves a number of purposes of which the most important, perhaps, consists in the calculation of probabilities. As will be noted in Chapter 16, a question that occurs and recurs in the field of mental measurement as well as in connection with most other scientific problems has to do with the likelihood that an obtained result merely represents a chance departure from zero. By ascertaining the number of standard deviations by which it exceeds zero (or any other specified point) and finding the *q* value corresponding to the *x* value in question, this probability is directly given if it is reasonable to suppose that the chances of variation are normally distributed as is generally the case. Thus if *x* = 2.326, *q* = .01, which means that there is only one chance in 100 that in other samples drawn from the same population the *direction* of the obtained finding would not be maintained. The *magnitude* of the difference, however, is likely to vary from one sample to another.

score method have attempted to provide normative standards for "the general population" or other vaguely defined universes, but as a rule they have tended to be more specific. Certainly nothing is gained by assuming, with no other evidence than is given by faith, that the mean and the standard deviation of scores on some newly devised test made by the six-year-olds in half a dozen New England towns correspond to those which would be found had all the six-year-olds in the United States been tested, to say nothing of those in other countries. There is far too much likelihood that with the passage of time faith may give place to hope and hope to a need for charity. One of the most encouraging indications of genuine progress in the field of tests and measurements is the increasing tendency to present standards for groups whose composition is described along as many axes as possible. Such expressions as "norms for ten-year-olds" or "the average fourteen-year-old" are by no means obsolete, but more people are asking how such standards were obtained and to whom they are supposed to apply.

OTHER WAYS OF EXPRESSING STANDARD SCORES: THE T-SCORE METHOD

In 1922, McCall noted the advantages of the standard score method both for the uniform expression of the results of various tests and measurements given to the same individual and for the calibration of test items in equally spaced units. After examining the statistics on school enrollment for a number of places, especially those for New York City, he came to the conclusion that a larger proportion of all living children were attending school at the age of twelve than at any earlier or later age, and that accordingly the data derived for twelve-year-old school children would be more nearly representative of the total population of that age than would those for any other single age group. McCall, however, failed to take account of the fact that not only the average score but also the variability of scores may change with age, even after the scores have been reduced to equally spaced units. Later investigations have in fact shown this to be the case. On the assumption of equal variability, McCall decided that the standard deviation of scores made by all the twelve-year-old children in the groups which he studied—chiefly those from New York City schools—would be a suitable unit by which to calibrate scales for the measurement of any characteristics which it was desired to study. Since twelve-year-olds were used as subjects for standardization, he called this procedure the *T-score method* and the unit of measurement a *T-score unit*. In order to avoid negative values and decimals, he placed the zero point of his scale at 5 S.D. below

the twelve-year mean and made the unit of measurement 0.1 S.D. He found that a range from +5 S.D. to -5 S.D. of the twelve-year distribution would include practically all cases likely to be found in the public schools, regardless of the test used.⁷ He therefore divided his scales into 100 T-score units, each unit representing 0.1 of a standard deviation of the twelve-year distribution. The mean score or normative standard for children of twelve years was thus automatically set at 50 with a standard deviation of 10; the standards for other ages had to be ascertained, but since they were always expressed in terms of the twelve-year distribution they could (if McCall's assumption of equal variability at all ages had proved to be correct) be directly compared with each other. Since this was found not to be the case, however, the T-score method as originally proposed gradually fell into disuse though it was exceedingly popular, especially in the field of educational measurement, until its insecure premises had become manifest.

By basing scores only upon homogeneous groups and making no attempt to give them more general significance, a more recent modification of the T-score method, which is still often called by that name, avoids the difficulty encountered by McCall. The advantage of this method is that it permits a comparison of an individual's performance on a large number of different tests with all scores expressed in similar units in which neither negative values nor decimals appear. The transmutation is easily performed. After finding the mean and standard deviation of scores for the group to which the subject belongs, the difference between the subject's score and the mean for the group is found and is then transmuted into S.D. units by dividing the difference just found by the S.D. of the group, carrying the division to the nearest tenth. The result is then added or subtracted, according to the direction of the difference, from 50 as a mean, as was done with McCall's twelve-year-olds. The formula thus becomes

$$\text{T-score} = 50 \pm 10 \frac{X}{\text{S.D.}_{dist.}}$$

where $\frac{X}{\text{S.D.}_{dist.}}$ is the difference between the individual's score and the average score of the group to which he belongs divided by the standard deviation of the distribution of the scores.

The advantage of this method is clearly seen in the following example. A certain college student is given a series of tests to help him

⁷ McCall was chiefly interested in measuring proficiency in the various subjects of the school curriculum and in tests of intellectual capacity designed for children of school age. He made no attempt to extend his measurements to the preschool level.

decide the kind of career for which he is best suited. His interests as well as certain extraneous factors point to one or another of three fields—journalism, business administration, or social service administration. Tests designed to determine aptitude for the three fields yield the following results in raw scores: journalism, 96; business administration, 54; social service administration, 116. Naïve consideration would at once assume that his aptitude for the field last mentioned is somewhat greater than that for the first and greatly superior to that for the second. But when the figures are reduced to terms of uniform meaning, the picture is very different, as is shown below:

	Mean Score (College Students) ^a	Subject's Score	Difference	S.D. _{dist.}	Diff. S.D. diff.	T-Score
Journalism	92	96	+4	20	+0.2	52
Business Adm.	40	54	+14	7	+2.0	70
Soc. Service Adm.	132	116	-16	32	-0.5	45

^a The norms are the means and standard deviations obtained for students of his college year (freshman, sophomore, etc.) enrolled in the same college or in other colleges of similar standards.

To the extent that the test scores are valid indicators of talent, one would have little hesitation in advising this young man to prepare himself for the field of business administration since his T-score on that test would be equaled or exceeded by only about one student out of fifty. Although his raw scores on the other two tests were much higher, when reduced to comparable units one was found to be only slightly above, the other somewhat below the college average.

Suppose that the average or normative score on all three tests had been the same, let us say 100 points measured in raw-score units, but that the standard deviation of the journalism test was 10, that of the test of business administration 25, and that of the test for social service administration 20. The subject's score on each of the three tests was 125. At first thought it might seem that with identity of the group averages with which he is to be compared and with identical scores on all three tests, his standing on all would also be identical. But a little consideration will show that this is not the case, since the wide variability of scores on the second and third tests means that there are a good many others who do as well as he, while the narrow spread of scores on the first places his score well out toward the extreme upper end of the distribution. The T-scores show this difference quantitatively. His T-score on the journalism test is 75; on the business administration test it is 60; and on the test of social service administration it is 62.5.

The widespread popularity of the intelligence quotient gives that method certain practical advantages over other devices for expressing the results of intelligence tests. Other quotients, such as educational quotients, social maturity quotients, and the like, which are calculated in a similar manner, have also come into general use. This fact has suggested the possibility of reducing standard scores to terms which have numerical values similar to those of intelligence quotients but which, because of the method by which they are derived, have fewer statistical limitations and hazards. Except for the use of different constants, the method is identical with the modern way of deriving T-scores. In place of assuming a constant mean of 50, the mean is set at 100 to correspond to the IQ of the average child. The question of a typical figure for IQ variability in terms of S.D. has never been settled. It varies with the test used and possibly with age. For the 1916 revision of the Stanford-Binet, which for many years was recognized as the standard intelligence test for children and adolescents, the variability of a representative group of American urban school children appears to be approximately 16-17 IQ points for ages six to twelve years. The variability of the 1937 revision is not the same at all ages. Although McNemar (1942) has presented a table for correcting IQ's on this scale to a constant variability, he does not state what variability was assumed as the correct one. From the data given in McNemar's monograph, together with the facts in the original manual of directions (Terman and Merrill, 1937), it appears that the typical variability on this scale may be slightly greater than that on the 1916 revision, perhaps 17 or 17.5 IQ points.

It should be noted, however, that if the T-score method (with different constants) is used for obtaining figures which correspond in general to IQ's obtained in the ordinary manner, it is no longer essential that the standard deviations of the mental-age equivalents increase proportionately to age or that different tests maintain the same variability in order that the T-score values may have uniform significance. All that is necessary is to *assume* a constant value for the standard deviation of the IQ's (or for that of other quotients, such as educational quotients) and to substitute this figure for the standard deviations actually obtained. It has been suggested (Goodenough and Maurer, 1942) that the term "IQ Equivalent" be used to indicate the results obtained by this method in order to avoid confusion with ordinary T-scores on the one hand or with IQ's obtained in the regular manner on the other. It will be noted that when this method is used, deviations are automatically made equal at all ages and accordingly (except for possible differences in the experimental error of measurement or of predictive value over a period of time) the significance of a given IQ is equal at all ages; that is, the

frequency of its occurrence is equal. The same rule, of course, holds for other quotients obtained in like manner.

The formula for obtaining the IQ Equivalent, assuming a constant S.D. of 17, is

$$\text{IQ Equivalent} = 100 \pm 17 \frac{X}{\text{S.D.}_{dist.}}$$

where $\frac{X}{\text{S.D.}_{dist.}}$ is equal to the difference between the subject's score and the average score of the group to which he belongs divided by the standard deviation of the distribution of scores made by that group.

The numerical significance of IQ Equivalents is comparable to that of IQ's computed in the ordinary manner. An IQ Equivalent of 125 indicates about the same degree of superiority, one of 75 about the same degree of inferiority, as we have been accustomed to attribute to corresponding levels of ordinary IQ's. However, certain basic differences between the two should be noted. In the first place, IQ Equivalents, unlike regular IQ's, are based entirely upon the performance of children of similar chronological ages and experience. In this respect they resemble percentiles rather than intelligence quotients. Although it is mathematically possible to estimate mental-age levels from IQ Equivalents by multiplying the child's chronological age by his IQ Equivalent (expressed as a decimal), it is questionable whether this practice is psychologically permissible. A young but very bright four-year-old child with an IQ Equivalent of 150, for example, is so unlike the child of six in many of his mental as well as his physiological characteristics that to imply mental similarity between the two is to misrepresent the facts. Better to note that only about one child in a thousand is as able as he. The use of mental ages is decidedly open to question when the mental level has been determined by the standard score method or by one or another of its derivatives such as the T-score method or that of the IQ Equivalent, where only a single age group has been used for transmuting scores into a more uniform system of notation.

A second point of difference has to do with the form of the distribution of the scores. The use of standard scores automatically converts the data into the form of a normal distribution. Although there is good evidence from many sources that for a reasonably representative group of subjects the IQ distribution is not far from normal, the true correspondence may not be as close as it inevitably becomes when the standard score method is used. It is unlikely, however, that the discrepancy, if it exists, is great enough to produce variations of much practical significance.

THE "DISCRIMINATIVE VALUE" METHOD
OF ARTHUR AND WOODROW

In 1919, Arthur and Woodrow described a method of scaling tests in terms of standard units, to which they gave the name of *units of discriminative value*. These units differ from ordinary standard scores in a number of respects. Arthur and Woodrow noted that in many tests, particularly form-board tests, picture puzzles, and others which are scored in terms of time, errors, or both, the amount of improvement from one age to the next is not uniform but is much greater at certain points of the age scale than at others. In such cases the magnitude of the standard deviations of the score distributions also undergoes a systematic change with age in most instances. These facts led to the belief that a unit of measurement which takes account only of the variability of a single age group is of questionable value, particularly in the case of a child at the border line between groupings. Moreover, a given deviation from the standard for his age would have a very different meaning if the child making the score in question were at a point in the growth curve where the differences from one age to the next are small and the standard deviations large, from that which would be assigned to the same amount of difference at another period when age changes are rapid and standard deviations small. These considerations led Arthur and Woodrow to use the average of the variabilities of two successive ages as the base of their measure^s and the difference between the mean scores made by two successive age groups as its numerator. The formula for the discriminative value of a given test at a given age is accordingly

$$\text{D.V.} = \frac{M_2 - M_1}{\frac{\text{P.E.}_1 + \text{P.E.}_2}{2}}.$$

A table of point scores is drawn up by setting an arbitrary zero point at a level just below the age at which success begins, computing the D.V.'s for each successive age difference thereafter, and adding them to build up standards for the later years. For example, in the Arthur Point Performance Scale (1930) for which this method of calibrating scores was used, the zero point is set at age five. The scale consists of a number of subtests for each of which discriminative values had been obtained. By interpolation, normative point scores based on D.V. values were

^s Arthur uses the probable error (.6745 S.D.) as the measure of variability instead of the standard deviation. This, of course, has no effect on the basic principle; it merely decreases the size of the denominator of the fraction and thereby increases its numerical value.

obtained for each month of chronological age within the range covered by the tests, and by extrapolation tentative norms for ages somewhat outside this range were computed. For example, on the Knox Cube Test⁹ the mean number of successes for the five-year-olds was 3.4 and for the six-year-olds, 4.9. The difference is 1.5. The probable errors of the two distributions were respectively 1.05 and 1.03, for which the mean is 1.04. The quotient or discriminative value obtained by dividing 1.5 by 1.04 is 1.44. Since the age groupings were counted to the last birthday, the average age of the children classed as five-year-olds would be 5.5 years; that of the six-year-olds 6.5 years; and so on. Therefore the discriminative value of 1.44 would apply to the age midway between these points, which is exactly six years.

In like manner, the D.V. for the interval between six and seven years was computed and found to be 1.33. This was added to 1.44 (since it is an additional credit), which makes the standard point score for age seven 2.77. In the same way, D.V.'s were computed and accumulated for each additional year up to age fifteen, after which little or no further gain appeared.

The same method was followed for the other tests used in the scale. Norms for the total were obtained by simply adding the point norms for the various tests at each age. The procedure is straightforward enough, and while some question may be raised as to the permissibility of averaging values which show as marked and consistent a change with age as do the P.E. values for many of these tests (see Arthur, 1933),¹⁰ the error introduced thereby is probably not very great. A more serious question arises with respect to the treatment of scores which are obviously of very unequal dependability as if they were equal. It is apparent from the data given by Arthur (1933) that at least in the case of her Point Scale, the relatively small average improvement from year to year at the older ages is accompanied by increased fluctuation of the P.E. values. This inevitably means less dependable determination of the D.V. values at the later ages. To some extent, this difficulty has been overcome by smoothing the curve to iron out some of the chance fluctuations in the

⁹ In this test four one-inch cubes are placed in a row at distances of one inch. With a fifth cube the examiner taps the cubes in a specified order, beginning with a very easy pattern such as 1234 (the numbers indicating the position of the cubes from the child's left to his right) and continuing to longer and more difficult patterns such as 13243, 143124 and 1321413. After each pattern has been tapped out, the cube is handed to the child, who is told to repeat the performance exactly. Only perfect repetitions are counted as correct; no partial credits are allowed.

¹⁰ For example, on the Seguin Form Board the P.E. drops fairly regularly from 7.15 seconds at age five to 1.17 at age twelve. On the Mare and Foal Test, the decrease is from 43.4 seconds at age five to 3.6 at age fifteen.

total, but the low self-correlations as well as the marked drop in the correlation of the test with the Binet for the later as compared to the earlier years suggest that a more basic difficulty is involved.

In this connection, it may be well to note once more a fact that has been mentioned repeatedly in this book. *When fine distinctions are called for, a more accurate measuring instrument is needed than is required when cruder distinctions will suffice.* In the case of such a scale as the one we have been considering, the changes in mean performance from one year to the next are very marked at the early ages but small at the later ages. If, owing to bias in the sampling of cases or to other causes, the obtained P.E. of a particular age distribution at the lower end of the scale is incorrectly determined, the numerator of the fraction will still be large enough (unless the error in the denominator is much greater than it is likely to be) to yield a D.V. unit of appreciable size. At the upper ages, an error of the same magnitude in the value of the P.E. is of greater significance since the small increase in average score with advancing age means greater overlapping of the age groups and requires more careful determination of standards if individual differences are to be accurately shown. This is especially true when, as in the scale under consideration, mental ages are to be computed on the basis of the D.V. total.

Other Devices for Interpreting Test Scores

In the preceding chapters the interpretative devices most frequently used have been discussed. There are, however, a few others which have aroused enough general interest to require brief mention.

THE HEINIS PERSONAL CONSTANT (PC)

In 1924 H. Heinis published an article in the French *Archives de psychologie* in which he applied a logarithmic transformation to the results obtained by Vermeylen (1922) from the application of a scale of mental tests, which he had devised to sixty children (ten at each age from six through eleven years) from the public schools of Paris. In spite of the small number of cases used, Heinis was particularly impressed by Vermeylen's study because the scale used included fifteen different kinds of tasks of ten items each, graduated in difficulty in such manner that each child's score would be a composite of his performance on each of the fifteen tasks. The time required for administering this scale was approximately three times that needed for a Binet test. Heinis believed that the added time indicated greater precision of measurement but, as Bradway and Hoffeditz (1937) have correctly pointed out, longer testing time does not of necessity mean greater accuracy in testing. The question is not how long a time is required to secure a sample but how well the sample represents the universe from which it is chosen.

Heinis's original interest lay in securing data from which the most representative curve of mental growth throughout the life span could be plotted. Vermeylen's results showed fairly regular increases in score at each age over the years from six to ten, but as he failed to present an account of the way the items were calibrated, the equality of the intervals from one score to the next was uncertain. He did state that a number of adjustments were found necessary on the basis of trial tests but did not describe the procedure used in making them. Furthermore, although he gave equal weight to each of the fifteen tasks, he presented no evidence that all were equally reliable and valid indicators of the trait they were intended to measure.

In his use of Vermeulen's data Heinis likewise failed to describe the details of his procedures. Inasmuch as neither the means nor the medians used for his statistical computations agree precisely with those calculated directly from the original data as reported by Vermeulen, it is apparent that some smoothing process was employed, but what that method was is not clear. On the basis of these corrected medians, Heinis concluded that the curve of mental growth is parabolic in form. Again without reporting his mathematical treatment in detail, he presented a formula

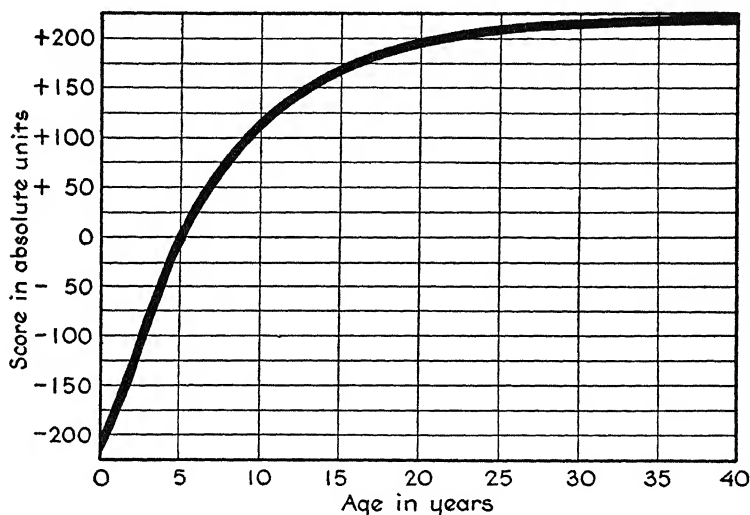


FIG. 12. THE CURVE OF MENTAL GROWTH AS CALCULATED BY HEINIS FROM DATA REPORTED BY VERMEULEN.

for the curve best fitting the results and by applying this to the ages above and below the range covered by Vermeulen's data he constructed the curve of mental growth shown in Figure 12, in which age is plotted along the abscissa and the ordinate indicates scores plotted in what Heinis believed to be equally spaced units. It will be noted that by this method the midpoint of the scale comes at about the age of five, which is later by approximately two years than that estimated by Thorndike (1926).

In a later article (1926), Heinis attempted to account for certain results reported by Kuhlmann (1921) on the re-examination of mentally defective persons by means of the Binet scale. When the results were expressed in terms of IQ's, it had been found that there was a fairly constant tendency for the latter to decrease with age. Kuhlmann also

found that the mental growth of idiots ceased at an age about three years earlier than that of borderline cases and that the rate of decline in the IQ increased somewhat with age. Heinis found that when a child's mental age was transmuted into the equally spaced units of the scale based on Vermeulen's data, and this transmuted score divided by the score (in Heinis's units) representing the standard for his chronological age, the resulting quotient to which Heinis gave the name of *the personal coefficient* (or *personal constant*) did not change with age as Kuhlmann had found the IQ to do. Heinis therefore decided that the PC gives a better prediction of later mental status than the IQ. Heinis thus defines the PC: "The personal coefficient of any given individual is equal to the result of the intelligence examination divided by the normal degree of intelligence corresponding to his age, both measures being given in absolute graduation."

For convenience, since the calculation of the PC is by no means a simple process, Heinis presented a table from which PC's could be read off directly after mental and chronological ages had been determined. Heinis apparently considered that this table could be used with any test for which age norms were available, but as we saw in Chapter 11, variations in standards from one test to another and for different age levels within the same test would introduce as many complicating factors into the PC as into the IQ.

Most of the published investigations in which the IQ and the PC are compared have to do with changes in the mean standing of groups after a period of time rather than with individual changes, and the greater number of these studies are based upon subjects of inferior intelligence, many of them being inmates of institutions for the feeble-minded. It is, of course, true that if, as a result of errors in standardization or from any other cause, there is a consistent tendency for an interpretative measure to change in numerical value with age, the probability is that an individual score would change in a corresponding manner. Accordingly if means and standard deviations at each age are known, the corresponding score of a subject at some later age would not necessarily be the same as that at an earlier age but could be calculated from the statistical facts.¹ If the interest centers around the stability of an

¹ Without allowance for regression, when the problem is simply that of determining what values at age *b* correspond to those at age *a*, the procedure would be like that of computing IQ Equivalents described on pp. 199-200. If, on the contrary, the question is one of making the best possible prediction of the later score of an individual on the basis of his earlier score, then the regression equation which takes account of the correlation between earlier and later tests must be used. The formula is

$$\bar{Y} = r_{xy} \frac{S.D._y}{S.D._x} (\bar{X} - M_x) + M_y$$

where \bar{Y} is the score to be predicted;

individual score rather than upon the tendency for group means to vary with age, Heinis's demonstration that when the PC instead of the IQ was used with Kuhlmann's feeble-minded subjects, no general tendency for the mean values to change with age was present does not answer the question at issue, since the individual values might still fluctuate more *in either direction* for the PC than for the IQ.

Several comparisons of the two interpretative measures have been made, but the results obtained are not entirely conclusive. Cattell (1933), using as subjects children from the Harvard Growth Study, found (again by a comparison of group rather than individual trends) that the PC varied less with age than did the IQ for those whose IQ's were less than 80 but that the opposite condition held for children whose IQ's were 120 or higher.

A direct comparison of changes in IQ and PC is not warranted, however, because the standard deviation of the latter is considerably less than that of the former. The difference increases as age advances. Table 5, which is based upon Hilden's tables of PC values (1933a), shows the corresponding figures for IQ and PC for mental-age deviations of 25 per cent above and below the chronological ages of four, eight, and twelve years.

Inspection of Table 5 shows that, particularly at the older ages, the PC values change but little even when the discrepancy between mental and chronological ages is as much as three years. While this inevitably makes for greater "constancy" of the PC values in terms of absolute magnitude, the gain in stability is accomplished only by the sacrifice of

r_{xy} is the correlation between first and second testing for the group to which the subject belongs;

$\frac{S.D._y}{S.D._x}$ is the ratio between the standard deviations of the two distributions;

\bar{X} is the subject's score on the first test;

M_x is the mean score of his group on the first test; and

M_y is the mean score of his group on the second test.

The logic of this equation is not difficult if one has grasped the principle of standard scores as given in Chapter 13. If the correlation between the two series of scores were perfect, then all that would be necessary for prediction would be to find the subject's standard score on the first test and then assign him an equal standard score on the second. To express the latter in terms of original units (raw scores) it would be necessary only to reverse the procedure used for converting original units to standard scores. If, however, the correlation is less than perfect, allowance for the lowered predictive value of the first test must be made by multiplying the number of standard units by which the first score exceeds or falls short of the mean by r and taking the result as the best indication of the subject's standard score on the second test, which can then be converted into original units of that test if desired. Note that allowance for regression gives a more conservative estimate of the subject's deviation from the mean, which is, of course, as it should be, since less can be predicted from an imperfect than from a perfect instrument.

TABLE 5
COMPARISON OF IQ AND PC VALUES
AT THREE AGE LEVELS

Method	Age 4	Age 8	Age 12
IQ	75	75	75
PC.....	80	85	89
IQ	125	125	125
PC.....	117	111	107

discriminative value. A measure which assigns all subjects pretty much the same rank is of little value for the study of individual differences. Moreover, the small number of cases upon which the Heinis scale was built, as well as the types of test used by Vermeylen,² would make it appear to be at best a lucky accident if the scale values should be found to hold for other groups or for other tests. It is therefore surprising to find that in the several investigations which have been made the PC has shown up as well as it has. The use of a logarithmic rather than a straight-line curve for expressing mental progress with advancing age is in line with most modern theories and research, but it is by no means certain that the same formula will hold good for all scales designed to measure mental growth. All in all, the superiority of the PC over the IQ has yet to be convincingly demonstrated.

THE PER CENT PLACEMENT

Like the other methods discussed in this chapter, the per cent placement has not been widely used in mental testing. Its chief protagonist at the present time is M. J. Van Wagenen, who has employed it as a means of expressing individual standing on a number of school achievement scales. The per cent placement must not be confused with the percentile rank described in Chapter 12. The latter, it will be recalled, is based upon the number of cases in a representative group of 100 of his mates whom the individual in question equals or surpasses in score. Since in the ordinary bell-shaped distribution, successive percentiles include equal areas above the base line of the curve, the distances which they subtend on the base line will become increasingly greater toward the extremes and shorter toward the midpoint. (See Figure 9.) In the case of the per cent placement, however, after equating the distances between score points by some kind of "absolute scaling" method, the

² An English translation of the Vermeylen tests has been published by Bradway and Hoffeditz (1937).

total distance between the positions theoretically occupied by the best and the poorest subjects in a group of 100 selected to constitute a representative sample of a specified age, grade, or other defined group is marked off into 100 equal divisions. These are numbered in order from lowest to highest, and the scores corresponding to each division point are arranged in tabular form. A subject is then said to have a per cent placement corresponding to the division point at which his score falls. In descriptive terms, we may thus say that the per cent placement indicates the per cent of the total distance between the lowest and the highest scores achieved by a representative sample of 100 cases of his age (or school grade or whatever other group is represented by the standard sample) which the individual in question has attained.

Like the percentile rank, the per cent placement is easy to understand. Unlike the former, it is measured in equally spaced linear distances and is therefore subject to arithmetical treatment. In both, the midpoint or average score is set at 50 rather than at 100, which makes it important to verify that all scores are properly labeled to avoid confusion with the IQ. The percentile method is the more popular since it may be applied without serious loss of accuracy to scores that have not been previously equated for difficulty, provided that they are arranged in correct serial order and the discrepancies in calibration, as judged from the form of their distribution, do not appear too great. The per cent placement, on the other hand, demands very careful calibration of the items if the standards are to be accurate.

THE MEDIAN MENTAL AGE

This method, which was first proposed by Pintner and Paterson, is a procedure for assigning a single representative mental age on the basis of a number of separate tests, each of which is scored separately and has its own set of mental-age standards. In 1917, Pintner and Paterson had developed a series of nonverbal tests, most of which were taken over directly or with slight modification from the work of Knox, Healy, and others. A certain amount of work had been done toward preparing normative standards for each of the fifteen tests in the series. In a monograph appearing in 1923 Pintner and Paterson presented more comprehensive norms and proposed several ways of combining the results of the entire series into a single score. Of these, the one which proved most popular is known as the median mental-age method. By the use of a table of standards for each test, the raw scores are converted into their mental-age equivalents. The median of this series of mental ages is then taken as the best approximation to the child's mental age that can be

had from the data. On the basis of his median mental age, the child's IQ can then be computed in the usual manner. This method is also used in certain group tests, notably the Kuhlmann-Anderson Group Intelligence Test.³

The median mental age provides a simple and reasonably accurate method of consolidating into a single figure the results obtained from a number of subtests designed to measure different aspects of the same function, particularly when the results of these tests are originally expressed in such diverse forms that some sort of conversion is demanded before they can be combined. For example, some of the tests in the Pintner-Paterson series are scored in terms of the time required to complete them, some in terms of the number of errors made, others in some system of points based on the degree of success. The numerical values of the scores differ so greatly as to means and variability that direct combination of them would be manifestly absurd. Reducing each to its mental-age equivalent and then taking the median of these mental ages gives equal weight to each of the subtests in the total.

A number of points may be noted here. The choice of the median instead of the mean mental age by Pintner and Paterson was based upon the observed fact that in their series of performance tests extreme and presumably nontypical scores were of fairly common occurrence. To allow these to exert their full influence upon the total would, so they felt, be likely to bias the results. It is questionable, however, whether the median is to be preferred to the mean in other scales in which this method of combining scores has been used, even though the choice seems to have been a wise one in the original instance. As has been pointed out in previous chapters, the mean is usually the more stable of the two measures, and the choice of a less stable in preference to a more stable device requires careful justification.

In the second place, the number of subtests used in determining the final score should not be too small. A decided convenience as well as a definite hazard of this method is the fact that some of the subtests may be omitted without producing any necessary change in the results except that of making them less dependable because of the smaller sample. Pintner and Paterson's complete scale included fifteen subtests from which the ten which proved most useful for their purpose were chosen to constitute what is known as their short scale. Ten tests are about as few as can be used to secure a median or even fair dependability unless the standard deviation of the mental ages is small because of high correlation between them—in which case, of course, the test will have a very

³ The Heinis Personal Constant has also been used as an alternative method of expressing results of this test.

constricted meaning. But it is not uncommon to find poorly trained clinicians who compute median mental ages from as few as four or five tests and draw conclusions from this meager evidence with all the assurance that comes from ignorance. The method is a valuable one for handling certain types of data when used with due appreciation of its requirements and limitations, but it should not be forced too far.⁴

OTHER INTERPRETATIVE MEASURES

Two of the earlier attempts to express test scores in terms that have uniform significance at all ages merit only brief mention since neither is often used at the present time. These are the "Coefficient of Intelligence" proposed by Yerkes, Bridges, and Hardwick for use in their Point Scale, and the "Index of Brightness" suggested by Otis. Both are based upon the direct numerical relationship between an individual or "raw" point score and the score which is the standard for the child's age without correction for unequal spacing of items or for the possibility of changing variability as age increases. The CI is the ratio of the point score made by a given subject to the score which is the standard for his age.⁵ The IB is found by taking the numerical difference between the subject's point score and the norm for his age and adding it to or subtracting it from 100 according to its sign. Although both these procedures yield results which have some superficial resemblance to the IQ, the sources of error in each should be too obvious to the reader to require further discussion here.

COMPARATIVE NOTE ON INTERPRETATIVE MEASURES

More than four decades have passed since Binet first proposed the mental-age method of transmuting into meaningful numerical terms a

⁴ The median may be preferred to the mean mental age in cases where a number of subjects of widely varying ability are given a single group test which may be well suited in difficulty to the majority but poorly adapted to those at the extremes, who are likely to make either zero or maximal scores on some of the subtests. Such scores do not, of course, indicate either zero or perfect ability but only that the subjects in question rank either below or above the limits of the test. How far above or below those limits they may be is not known. Their scores are therefore indeterminate and as such cannot justifiably be used in reckoning a mean, but to fail to take any account of them would be an even greater error. They are accordingly entered as "below MA—" or "above MA—" (the mental levels corresponding to the lowest and highest possible scores on the test in question) and if the number of such scores is not too great, a median mental age may be computed without serious loss in accuracy, since the exact value of the extreme scores is unimportant if their position in the series is known. (See Chapter 12.)

⁵ Had this scale been calibrated in equal units of growth, the CI would have been identical with the Heine PC.

qualitative or semiquantitative description of a child's performance on a series of mental tasks. Since then psychologists and statisticians have experimented not only with this method but with many others as well. Of these the intelligence quotient which is derived from the mental age continues to outrank all others as far as its use with children is concerned, but the standard score method in one or another of its variations is steadily gaining in popularity, especially among research workers. Standard deviation units have now become the almost universal basis for constructing and calibrating scales designed for use with mature subjects for whom advance in age is no longer a useful measuring rod. Even in the case of children, the advantages of limiting comparisons to those whose physiological development and life experience are at least reasonably similar to each other are becoming more clearly recognized. Although the method of percentile ranks has this merit in common with standard scores, the fact that the latter are equally spaced along the base line of the distribution curve, while percentiles are not, lends to the standard scores a distinct advantage, especially for purposes of research.

Although the use of standard scores goes back to the early days of testing in the United States, the popularity of the method is comparatively recent. Other terms are more readily understood by the layman, and mental inertia has always been a stumbling block in the way of progress. Now that more people are becoming acquainted with the advantages of the standard score method, it is to be expected that more scales will be standardized on this basis and that more uniform devices for making the results directly meaningful to persons without statistical training will be adopted.

Testing the Tests. I. General Principles and Fundamental Methods

CHECKS TO BE APPLIED

In the literature of mental testing two terms which have come into general use call for some explanation and comment. These terms are "reliability" and "validity." *Reliability* is usually defined as *the degree to which a test is consistent in measuring whatever it does measure*. By *validity* is meant *the accuracy with which a test measures what it purports to measure*. The distinction is an important one, and the concepts underlying the definitions have played a significant part in the history of mental testing. As not infrequently happens, however, the widespread use of the two expressions has in many cases led to stereotyped thinking, to the substitution of formulated procedures for intelligent consideration of the requirements of the problem at hand, and to dogmatic assertions on the basis of statistical findings "interpreted" by rule of thumb—and frequently a very clumsy thumb.

The term "reliability" in particular seems ill chosen. In everyday speech we apply the term "reliable" to a person or thing regarded as trustworthy, dependable, accurate. All these terms have an implied reference to the requirements of some particular situation. It needs but little questioning to demonstrate that a goodly proportion of those who use the term in its statistical sense have by no means freed themselves from an unconscious feeling that a statement about the "reliability" of a test has some bearing upon its dependability for its designated purpose. Verbal habit is stronger than most of us realize. For this reason we have preferred to use the term "stability" when a general reference to consistency of results is called for and to specify the factors correlated when citing the results of particular investigations.

The practice of designating coefficients of correlation as "reliability coefficients" regardless alike of the factors correlated¹ or of the charac-

¹ Three methods of obtaining "reliability coefficients" are in common use: (1) the correlation between the scores earned on an initial test and a retest of the same

teristics of the group whose scores are correlated is much to be deprecated. Although the effect of increasing the range of talent upon the correlation coefficient is—or should be—familiar to every student of elementary statistics (see page 163 ff.), it is still unfortunately common to hear persons who should know better stating that “the reliability of a certain test in thus and so” as if the figure cited were a fixed value, independent of the specific conditions under which it was derived. Actually, the determination of a coefficient of correlation should be regarded as an initial rather than a final step in ascertaining the stability of test scores. Kelley (1923) has pointed out that if a test is equally effective throughout the range of talent covered, the standard error of estimate of a true score² will be unaffected by group variability. This figure, therefore, has a stability that is lacking in uncorrected coefficients of self-correlation.

For example, suppose that the correlations between test and retest as reported in the literature for two different scales purporting to measure the same function were respectively $+.90$ and $+.60$. Naïve consideration might well lead to the conclusion that the first is more “reliable” than the second. But let us look into the matter a bit further. We note that the first correlation was based upon the scores made by subjects with an age range from eight to sixteen years; in the second case all were between the ages of ten and eleven years. The standard deviations of the scores on the first administration of the tests were 24 and 6 points, respectively. Examination of the age norms reported for the two tests (each stated in terms of its own scale values) shows that the standards for ages nine to eleven years are as follows:

Test	Ages		
	9	10	11
A	32	38	42
B	20	28	37

individuals by means of the same scale after a time interval so short that no real change in standing is likely to have occurred; (2) the correlation between scores earned on the two halves of a scale divided according to some systematic plan (such as the sum of the scores on the odd-numbered versus that on the even-numbered items) after correction for the total length has been made by means of the Spearman-Brown prophecy formula; and (3) the correlation between two presumably equivalent forms of the same scale. Although the term “reliability coefficient” is by many people applied indiscriminately to the results of all three methods, these results are unlikely to be equal, nor can they be looked upon as having the same psychological meaning. (Jordan, 1935; Goodenough, 1936.)

² See footnote on p. 166.

The standard error of estimate of a true score on Test A is $24\sqrt{.90 - .90^2} = 7.2$; on Test B it is $6\sqrt{.60 - .60^2} = 2.9$. Thus, in spite of the greater magnitude of the "reliability coefficient" of the first test, the standard error of estimating the true score of a given child on this test is more than the typical increase in score from one year to the next, while in the case of the second test the error of estimate is only about a third of a year's growth. It is apparent that the person who places his trust in uncorrected self-correlations as reported in the literature is likely to be woefully misled in his search for a test that will yield stable results.

THE CRITERION MEASURE

No matter how stable the results of a test may prove to be, the method is of little practical use unless the meaning of the scores in terms of some larger or more generalized context is known. And if a test is needed for some particular purpose, the fact that it is a dependable measure of something else does not help much. It is usually possible to improve the stability of a measure which has shown itself to be a valid indicator of the characteristic it is desired to appraise, while only a radical change in procedure will serve if the method proves to be unsuitable for the purpose at hand. The primary concern, then, both of the test maker and of the test user, must be the extent to which the test results agree with some accepted criterion measure. In the absence of such a criterion, progress will necessarily be blind and uncertain, for there will be no way of checking its direction. For this reason the first efforts of the test maker should be directed toward securing data which may be used as a criterion by which he can guide his progress. For like reasons the test user should also have some objective way of determining how well the tests that he administers are serving his purposes. This implies that he must have a clear and unambiguous understanding of what these purposes may be.

What is needed, then, is something in the nature of a criterion for the criterion. If taken literally, however, this might seem to call for an endless chain in which one criterion demands another after the manner of the well-known couplet:

Big fleas have little fleas upon their backs to bite 'em
And little fleas have smaller fleas and so *ad infinitum*.

There are several possible ways of resolving the dilemma. One way is to narrow the field designated by the test. For example, a group test may be developed as a less expensive substitute for an individual test. Now the individual test may not be wholly valid when looked upon as a

measure of that which it purports to measure. But the scores obtained by its use are objective facts, the stability of which may be determined and defined by one or another of the methods previously described. If the new group test is frankly called "a test for predicting scores on the ——— test of ——— ability," a criterion is thereby chosen about the nature of which there can be no possible misunderstanding. If the self-correlations of both the criterion measure (the individual test) and the group test are ascertained, *using the same group of subjects in both instances*, Spearman's formula for the correction of "attenuation" in the correlation between two related measures resulting from the experimental errors in each³ may be applied to ascertain what the theoretical correlation between the two tests would presumably be if both were completely freed from errors of measurement.

Properly interpreted, the correlation between a test and its criterion after correction for the attenuation factor can be very informative. An example or two will illustrate. Suppose that an individual test which is taken as the criterion measure is found to have a self-correlation for a given group of subjects of $+.81$. For the same subjects a group test which is designed as a substitute for the individual test has a self-correlation of $+.90$. If the only factors making for lack of agreement between the two sets of scores were variable errors of measurement, their correlation should be $(\sqrt{.81}\sqrt{.90}) = +.85$. Actually it is found to be only $+.40$. This at once shows that there are fundamental differences between the two tests which cannot be corrected without fairly drastic changes in the content or the procedure of the new test. Merely increasing its precision without changing its character will not serve. If, however, the obtained figure had been close to the possible maximum, say $+.83$ or $+.84$, one would be justified in assuming that the new test is not only an acceptable

³ It is apparent that the results obtained by the use of an inaccurate measuring instrument cannot be expected to show exact correspondence to anything. It can be shown that the highest possible correlation between a perfectly accurate series of measurements of a given function and the scores obtained by the use of a somewhat unstable measure of the same function is the square root of the self-correlation of the latter. If both measures are unstable, but apart from their experimental errors both measure the same thing, their maximum correlation with each other will be the product of the square roots of their respective self-correlations. The ratio of this maximum correlation to that actually obtained thus shows the extent to which the two really do measure the same thing insofar as they measure anything at all. The formula is

$$r_{cor. \text{ for at. }} = \frac{r_{12}}{\sqrt{r_{11}} \sqrt{r_{22}}}$$

where r_{12} is the obtained correlation between the two variables and r_{11} and r_{22} are their self-correlations. It may be well to repeat here, since the principle is so often disregarded, that the same group of subjects must be used throughout if the attenuation formula is to hold.

substitute for the criterion measure but may even be somewhat superior to it, since the self-correlation is higher.

A word of caution with respect to the use and interpretation of correlations corrected for attenuation is needed. First, the groups upon which these correlations are based must be large enough to ensure that the magnitude of the coefficients is dependably established. Second, it is necessary to make sure that the self-correlations have not been vitiated either by correlation between errors which would tend to make them too large, or by lack of correspondence between the measures correlated (parallel forms, split halves, or repetitions of the same form) which would tend to make them too small. Correlation between errors is particularly likely to occur when ratings by two judges take the place of more objective measures, especially if there has been opportunity for discussion. Even when no such possibility exists, there is still the likelihood that both judges may have been influenced in their ratings by the same kind of extraneous factors such as failure to take account of age differences among children in the same grade when estimating their intellectual levels, or considering such matters as race, political affiliations, and the like when rating the efficiency of workmen. When a test is repeated, the comparability of the two trials may be affected by the tendency of some subjects to look up the correct answers to remembered items or that of others to give identical responses on the two occasions merely as a result of memory. Habituation to the test situation as such, leading to greater ease and self-confidence on the second occasion, must also be taken into account. When parallel forms of the same test are used to determine the self-correlation, not only may the two presumably comparable forms be psychologically dissimilar, but the results may also be affected by differences in the physical condition or emotional attitudes of the subjects on the two occasions. Unless great care is taken to secure stable and unbiased values for the self-correlations, the use of the correction for attenuation may be highly misleading.

The advantages of designating a test in terms of the criterion used rather than by some general and not too clearly defined trait name, such as "emotional stability," "mechanical ability," or "honesty" are apparent. One cannot ascertain "how well a test measures what it purports to measure" without knowing rather precisely what that purport is. Far too many test makers have designated their tests by names that are much more ambitious than is warranted by the evidence they present. For example, in the illustration just given, the group test might have been called "A group test of _____ ability." Had this been done, however, the individual test could no longer be regarded as a wholly satisfactory criterion measure. It is merely another approach to the same problem.

Under the title first suggested, the universe to be sampled has definite boundaries which are set by the individual test. Under the second title, the boundaries are indeterminate. If the group test, in spite of high correlation with the individual test, fails to meet the requirements of some practical situation, the first title immediately suggests the most likely source of difficulty. But if, as is often true, the nature of the criterion actually used in standardizing a new test is concealed under some more ambitious title, sources of error are less clearly marked. The way to test improvement often lies along the road of more modest claims.

Test scores are not the only kinds of objective criteria in terms of which a new test may be standardized and designated. The occurrence of certain events or forms of behavior may be the composite results of a wide variety of factors, but the events themselves are beyond question. For example, a penitentiary sentence undoubtedly depends on a host of conditions in addition to chance factors, some of which pertain to the individual, others to his surroundings, and still others to the relation between the individual and his environment. In spite of all this, it might well prove more fruitful to use such a criterion, either in an all-or-none form or as a continuous variable in terms of the length of the sentence imposed, for the development of a measure designed to identify prospective criminals rather than to struggle along with attempts to define and measure "criminal tendencies" or "the delinquent personality."

Hull (1928) has pointed out that the criteria commonly employed for the validation of mental tests are of three kinds: (1) material products which can be measured or appraised in some objective manner, (2) activity which can be observed and recorded at the time of its occurrence, and (3) subjective judgments by persons presumed to be competent for the task. The first two have the great merit of objectivity, and if the conditions under which the data are secured are kept the same for all, and these conditions are carefully described so that the exact nature of the criterion is known to all who use the measure, it is not only possible to ascertain what proportion of the variance in the criterion measure is accounted for by the variance in the test, but also, if desired, what part of the test variance does *not* coincide with the variance in the criterion, i.e., fails to contribute to the purpose for which the test was designed. In this way may be determined both the insufficiencies arising from failure of the test to account for the total variance in the criterion and the wastage resulting from the fact that a part of the test variance will usually be found to result from factors other than those responsible for the criterion variance. This procedure, which is known as an analysis of the variance, was developed by R. A. Fisher in a long series of brilliant investigations relating, for the most part, to the field of agriculture. The

application of Fisher's methods to mental testing is more recent but is rapidly gaining in popularity. A brief account of the method will be presented in Chapter 18; a discussion of its more refined aspects would fall outside the scope of this book. Snedecor (1946) has described Fisher's methods in detail, although most of his examples are taken from agricultural research. Lindquist (1940) has shown how the procedures may be used in the study of educational problems. Less detailed accounts may also be found in practically all the more recent textbooks on statistical methods.

The methods of multiple and partial correlation are useful in determining how well the separate elements of a composite measure or a series of related measures predict the scores on a criterion measure. Since it is usual to find that the variances of the part scores duplicate each other to some extent, and also that each part involves a certain amount of wastage, the problem becomes that of weighting each part in accordance with its unique contribution to the total. This procedure is known as multiple correlation. An excellent account of the method and of its underlying theories has been given by Guilford (1936).

The multiple-correlation coefficient represents the correlation between a criterion and the sum of a series of measures used for predicting it when each individual measure is given its best possible weight in the composite. It may therefore be regarded as a synthetizing device whereby each of the available measures is made to yield its maximum contribution to the purpose at hand.

Partial correlation is, in a sense, the opposite of multiple correlation. It is a device for analyzing an obtained correlation into more nearly homogeneous elements by ascertaining what its magnitude would be if it were freed from the effect of one or more of its determinants by rendering all the scores on that variable constant or nearly so. For example, if children whose ages range from four to ten years are used to determine the correlation between height and weight, the obtained r would be very high, but much of the apparent relationship would be the result of the uncontrolled variance in age. In this case, partial correlation enables one to determine the most probable value of the correlation between height and weight among children whose ages differ only by some specified smaller amount, say by not more than six months, or any other range that the investigator wishes to use.⁴ Although a positive relationship as a rule will still be found, the reduction in the variance that is due to the

⁴ It will be noted that in the case of continuous variables the statement that scores in the homogenized variable have been "rendered constant" is inexact. As a rule, a grouping factor will still remain although the groups may have such a narrow range that the variance within each is so small as to be negligible.

age factor will bring about a corresponding reduction in the partial correlation. A correlation of which the magnitude has been changed by the statistical elimination of all or most of the variance due to a single one of the factors by which it was determined is known as a *first-order partial correlation*. *Second-order partials* may be found in the same way as was done in the first instance (using the first-order partials as a base) in order to ascertain the degree of relationship to be expected if two of the contributing factors are made approximately constant. If desired the procedure may be continued until all of the measured factors, except the two for which the extent of the basic relationship is in question, have been rendered constant in turn. However, unless the number of cases is large, the error of estimate involved in this continuous process of making one approximation from another becomes increasingly hazardous. Moreover, the amount of computation increases very rapidly as new variables are added to the series. For these reasons it is not usual to find much work done with partials of an order higher than the second.

The multiple-regression equation is sometimes useful in developing a criterion which cannot at the time be measured directly. Suppose, for example, that one were attempting to work out a new test for the prediction of success in some kind of industrial occupation. Success cannot be determined directly without trial, which is costly if many of the candidates prove inefficient. If, therefore, a method of predicting success which can be used at the time of application could be developed, the firm could be saved a good deal of unprofitable expense. Provided that experience on the job does not affect the test scores or that it does so to an approximately uniform degree for all subjects, experienced workers could be used as a criterion group by which the value of a battery of tests designed for use with candidates for new positions could be determined. In such cases it not infrequently happens that a number of relatively short tests are found to show low but dependably positive relationships to the criterion; each one accounting for a part of the variance in the latter but not for a sufficient amount to make it a valid indicator if used alone. By the use of multiple correlation, weights for each of these tests can be found. On the basis of this information a test battery can be devised with each of the parts given its appropriate weighting. The scores from such a test may be used directly for prediction or as a criterion from which other and possibly still more dependable tests can be developed.⁵

⁵ The method of successive approximations in test building has not been developed statistically as far as would be desirable. Much of the success of the Binet tests is undoubtedly due to the large number of times these tests have been revised when sources of error in the older formulations were discovered. For the most part, however, the changes in the successive revisions have been made on the basis of

If a series of measurements of the criterion group is available, the extent to which variance in the selected criterion measure is the result of variance in one or more of the other measures of the criterion group can be determined either by the methods of partial correlation or more directly by analysis of the variance. For example, if the efficiency of a group of women office workers were to be determined solely on the basis of the judgments of a susceptible male supervisor, a good deal of later trouble might be saved by securing judgments of their personal attractiveness—preferably by the same supervisor—and rendering this factor constant by means of the partial-regression equation⁶ before using the ratings as a criterion for a more objective test.

Even when a criterion is supposed to be entirely objective, it may still be found that errors are not distributed at random but show a constant trend. Some years ago I had occasion to make use of the scores on a spelling test which had been administered and scored by the classroom teacher. Because some of the scores were unexpectedly high or low, I asked to see the original papers. Many errors in scoring were found, as well as a number of cases in which the spelling was uncertain because of illegible handwriting. The teacher was then asked to arrange the children in rank order according to her judgment of their "attractiveness of personality." The correlation between the number and direction of the errors in marking the spelling papers and her judgments of the children's personality was $+ .42$. Errors made by the children whom the teacher thought most attractive were likely to be overlooked and illegible spellings marked "correct," while the youngsters deemed unattractive were not only rarely given the benefit of the doubt if a word was not clearly written but in some cases had also been unjustly penalized for words correctly spelled. Such terms as "chance," "experimental errors,"

observation and clinical judgment combined with rather haphazard use of statistical methods. More consistent and considered use of such methods as multiple correlation, analysis of the variance, and the like in which new criterion groups are used from year to year, new tests are added, old ones eliminated, and weights changed would in all probability show that more can be accomplished through successive revisions of old tests when done on an objective and theoretically sound and consistent method than by continued attempts at developing wholly new procedures or by revising earlier tests on the basis of incompletely verified "hunches" and a few statistical results secured without consistent theory or plan.

⁶ As noted elsewhere, the regression equation provides a method for estimating the most probable value of the scores on one variable when those on a second related variable, together with the magnitude of the correlation between the two and their standard deviations, are known. The partial-regression equation makes it possible to estimate the most probable value of the scores on some measured characteristic if those on some other related characteristic were all alike. In the example above, it would enable one to estimate the efficiency ratings of the workers that would be given by the supervisor if he regarded all as equally attractive.

"errors of measurement," and so on are convenient terms under which a good many carelessly made assumptions are frequently cloaked. Actually, of course, every consequent has its antecedent and that portion of a series of test results which we call "chance" merely represents those consequents for which the antecedents have not been ascertained. It should be the concern of every test maker to bring to light as many of these unknown factors as possible, even if he cannot prevent their occurrence. In like manner the test user who has neither the time nor the facilities to improve the tests he uses should still be on the alert to ascertain, as far as is in his power, where their deficiencies lie and how he can best make allowances for them. Often a relatively small amount of carefully planned statistical examination of the tests already available will provide more useful information than would an equal amount of time spent in attempts to develop new methods.

THE RELATIVE IMPORTANCE OF THE STABILITY OF TEST RESULTS AND OF THE VALIDITY OF THE CONVENTIONAL INTERPRETATIONS OF TEST MEANING

It was pointed out in an earlier paragraph that "validity" in the usual sense of the term should take precedence over "reliability." This is equivalent to saying that if one is in need of a tool for a particular purpose, even a comparatively crude instrument that will to some extent serve that purpose is to be preferred to a more perfect one that is useful only for something else.

A corollary to this fact is of particular interest to the constructor of tests. Suppose that in the course of his researches he succeeds in developing two quite different measures, both designed for the same purpose and each having a correlation with a given criterion measure of $+.40$ for the same group of subjects. In both cases the time required for the test is short enough to allow at least a fourfold increase without becoming excessive. The first test has a self-correlation of $+.90$; the second, one of $+.50$. The criterion measure has a self-correlation of $+.80$. Which is the more promising test?

If nothing further could be done with either, then there is no question but that the test which yields the more stable results and is as valid as the other is the one to be preferred. But it was noted that the time requirements are such that the sample of items included in each could be quadrupled, if necessary. Since a large sample gives a more dependable estimate than a small one, it is to be expected that lengthening a test would increase its dependability as a measure of whatever it

does measure. Since it has been shown that each of the tests in question does account for a certain amount of the variability of the criterion, it is likewise to be expected that, because of the elimination of some of the chance factors, an improvement in the stability of their results would also improve their validity for the purpose at hand.

In 1911, Brown, working on the basis of a principle previously developed by Spearman, presented a formula for predicting the increase in the magnitude of the self-correlation of a measure to be had by increasing the size of the sample. His formula, generally known as the Spearman-Brown prophecy formula, is given below:

$$r_{af\ af} = \frac{ar_{11}}{1 + (a - 1)r_{11}}$$

where $r_{af\ af}$ is the predicted self-correlation of a sample which is a times as large as the original one; and

a is the proportional number of times the size of the sample has been increased.

In the example given, we note that a fourfold increase in length might be expected to raise the self-correlation of the first test to $+.97$; that of the second, to $+.80$.

What about the effect on the correlation with the criterion? This can be estimated from the following formula (Kelley, 1923):

$$r_{c(af)} = \sqrt{\frac{1 - r_{11}}{a}} + r_{11}$$

where $r_{c(af)}$ is the predicted correlation with the criterion of a sample a times as large as the one previously secured;

r_c is the correlation with the criterion of the original sample; and

r_{11} is the self-correlation of the original sample.

Application of this formula to the data of our problem shows that the first test cannot be greatly improved by securing a larger sample of the same kind, since its predicted correlation with the criterion resulting from a fourfold increase in length is raised only from $+.40$ to $+.43$. For the second test, however, the correlation is changed from $+.40$ to $+.64$. The reason for the difference lies in the fact that in the one case only a small part of the difference between test and criterion could be accounted for by instability of the scores obtained by a test which already had a self-correlation of $+.90$. Accordingly, the mere securing of more data of the same kind could not be expected to help much. In the second case, however, the low self-correlation, coupled with the fact that, in spite of the instability of the test scores, the correlation with the criterion was as high as that of the first test, suggests that a somewhat greater part

of the lack of agreement between test and criterion resulted from chance variations in the former which were in part corrected by increasing the size of the sample.

The use of the two formulas just given presupposes (1) that the added portions of the test are basically so similar to the original part that all intercorrelations and standard deviations are equal, and, (2) that testing conditions are such that it is practically feasible to devote the necessary time to obtaining the larger sample. One occasionally finds instances in which these formulas have been used when one or both of these conditions are manifestly out of the question. Thus we find makers of rating scales for use with school children reporting on the correlation between the ratings of a group of children by the previous and the present classroom teachers. They then report what the correlation would probably become if ratings were secured from six, ten, or some other hypothetical number of teachers, ignoring alike the improbability of finding that number whose acquaintance with the subjects would be sufficient to enable them to make valid judgments, and the further fact that differences in standards and in the ways of regarding people and their behavior are such as to make it very unlikely that the ratings given by one teacher would be entirely comparable to those given by another. In other instances the maker of the test may have pretty well exhausted his ideas when drawing up his original series of items. Other forms, if devised, would have to be made up largely of the leftovers and would therefore be unlikely to be as useful as those in the original series. The use of the Spearman-Brown formula is thus very questionable except in those cases where (a) comparable forms have been devised in advance which have been shown to meet the conditions of equal standard deviations and intercorrelation, or (b) the test is of such a simple and uniform nature as to make it obvious that it can be lengthened without changing its essential character. The latter condition is exemplified by tests of arithmetic computation such as single-column addition problems, or of speed of tapping as measured by the number of taps made in a given short period of time. It is, of course, assumed that sufficient rest periods are interposed between the successive trials to avoid the likelihood of incomparability resulting from fatigue.

If the necessary conditions are met, however, both the test maker and the test user will find the prophecy formulas valuable guides to further procedure, since they enable one to predict with considerable accuracy how much is to be gained by increasing the size of a test sample or, conversely, by how much a test must be lengthened in order to attain a required degree of stability or a given correlation with some chosen criterion. They permit economy of time by making it possible to do much

of the spadework of trying out tests with relatively short forms in order to find which methods are most promising for the purpose at hand. By substituting small-scale experiments for long and costly ones, more efficient use of both time and funds is made possible.

MULTIPLE CRITERIA

It not infrequently happens that no single criterion is to be had which can be regarded as an unbiased indicator of the characteristic a test is designed to measure, but that there are a number of available measures which may reasonably be looked upon as partially related to it. In such cases a study of the correlations of the test with these partial criterion measures is often informative. Inasmuch as the criteria are likely to differ with respect to their self-correlations, the correction for attenuation should be applied in order to show the relative degree of relationship of the test to the different criteria apart from errors of measurement.

If the various partial criteria are of diverse kinds, multiple correlation may be used to show the extent of their combined relationship to the test; in other words, the extent to which the test variance can be accounted for in terms of these partial factors taken in combination. It should be unnecessary to point out that no assumptions as to *causation* are warranted on the basis of such a procedure. Correlation, whether simple or multiple, merely indicates the existence of common factors within the measures correlated.

FACTOR ANALYSIS AND THE "PURIFICATION" OF TESTS

Thurstone's recent volume (1947) sets forth in an admirable fashion the underlying theory of factorial analysis in its relation to test construction and interpretation. No adequate account either of this book or of the methods which it describes can possibly be given here; we shall note only the general principle upon which all factorial methods, despite differences in details, are based. In general, the position taken is that what appears to be an almost infinite variety of forms and levels of behavior may be only the permutations and combinations of a relatively small number of underlying factors or directional tendencies which, taken in their totality, describe the individual in all his uniqueness. Because any act (such, for example, as responding to a mental-test item) is the result of a number of these factors acting in combination, mental factors cannot be isolated and recognized by surface inspection nor measured in terms of superficial assumptions as to how and where they

are manifested. But if a comparatively large number of measurements of related forms or behavior or types of ability are made, such, for example, as would be given by a series of different kinds of intelligence tests or tests involving memory or mechanical skill, and the intercorrelations of these measures for a particular group of subjects are worked out and arranged in the usual form of rows and columns so as to produce what is known as a *correlation matrix*, a certain orderliness in the numerical relationships of homologous terms can be noted. As early as 1904, Spearman called attention to the fact that if the intercorrelations of any four tests are chosen from such a matrix, the products of two homologically situated diagonal values will be equal (within the limits of chance variation) if a single factor accounts for the intercorrelations. Thus the difference between the "cross products," as they are called, becomes approximately zero. By rearrangement of the table, three such *tetrads* or groups of four can be formed from the intercorrelations of any set of four tests. Thus a test can be had of the stability of the tetrad-difference criterion. These tetrads, where the subscripts *a*, *b*, *c*, and *d* indicate the four tests, are shown below:

$$t_{abcd} = r_{ab}r_{cd} - r_{ac}r_{bd}$$

$$t_{abdc} = r_{ab}r_{cd} - r_{ad}r_{bc}$$

$$t_{acdb} = r_{ac}r_{bd} - r_{ad}r_{bc}$$

when *t* is the tetrad difference.

If only one common factor accounts for the correlation among the four variables, *t* will be zero,⁷ since the variance not common to all four has been removed by subtracting the cross products from each other, thus leaving only the common factor in each term. This was so generally found to be true in the case of Spearman's early studies that it formed the basis for his famous two-factor theory of intelligence. (See Chapter 19.) However, the tetrad difference method is no longer used to any extent; it has been supplanted by more efficient procedures in which the underlying assumptions are less dogmatic and more in accordance with common observation. The person chiefly responsible for the methods in current use today is L. L. Thurstone of the University of Chicago, whose recent monograph (1947) describes these procedures both as to their mathematical derivation and their psychological setting and interpretation. Thurstone calls attention to the fact (as many others had done before him) that while the tetrad difference criterion is satisfied in many instances, its divergence from zero is so often found to be greater than could reasonably be accounted for by chance that the assumption that all

⁷ Within the limits of chance variation.

intellectual differences can be accounted for on the basis of a single common factor plus specific factors, as Spearman believed to be the case, becomes untenable. (See Chapter 19.)

Thurstone's procedures are based upon these assumptions: first, that mental organization is not haphazard but possesses a certain orderly structure, the nature of which can be shown, at least to a first approximation, by statistical treatment of the intercorrelations found for a series of tests; second, that this structure can be roughly described in terms of certain underlying abilities or characteristics known as *factors*, which differ from those measured by the tests in being more nearly simple and homogeneous. The variance in each test usually is made up of a composite of the total number of factors tapped by the entire series of tests. The factors, however, are not all of equal importance, since some account for a greater proportion of the total variance than others do. In carrying out a factor analysis by Thurstone's methods, one isolates first the factor accounting for the greatest proportion of the total variance. This is known as the *first factor*; it probably corresponds fairly closely to Spearman's general factor though arrived at by somewhat different methods.⁸ The portion of the variance not accounted for by the first factor is known as the *first residual*, and unless this residual is small enough to be accounted for by chance it is apparent that other factors have played a part in determining at least some of the intercorrelations. A second factor is then extracted and the residuals are calculated as before. The process is continued until the residuals have, for the most part, been reduced to a point at which their magnitude does not reliably exceed that of the average r of the original correlation matrix.

Factorial analysis of a battery of tests or of the items comprised within a given test makes it possible (1) to determine whether or not a given test or test battery measures a single unitary characteristic or a complex into which a number of different factors enter in varying degree; (2) to ascertain what is the smallest number of factors that must be postulated in order to account for the intercorrelations and (3) what proportion of the total variance of each item or test in the battery is accounted for by each of the factors underlying the total matrix; and (4) by means of regression equations to estimate, from his scores on the tests that depend on these factors, the probable standing of an individual on each of the primary factors underlying the correlation matrix.

The first point is of particular importance when dealing with tests that are designed to measure some general but presumably unitary factor.

⁸ It should be noted that Spearman's original methods have since been considerably modified by him and his followers although the two-factor theory itself was not essentially changed.

If the variance of the different items or subtests is really determined by a variety of different unitary factors rather than by the single homogeneous one that is postulated by the name given to the test, interpretation of scores becomes difficult unless the nature of the factors and the relative weight of each in the different parts of the tests are known. McNemar (1942) made a factor analysis of the items of the 1937 revision of the Stanford-Binet which showed that a single factor could account for most of the variance throughout the items of this test; in other words, that while it could not be stated with certainty that only one mental characteristic completely determined the scores on this test (apart, of course from experimental errors of measurement and variations in the sampling of subjects), additional factors played at most a very minor part. Most tests, however, have been found to depend on more than one factor. Particularly is this true of group tests made up of a number of subtests each of which comprises a fairly large number of items of the same kind. It is very likely that the fact that these tests usually show lower correlations between measurements of the same subjects after an interval of time is in part due to the nonunitary character of the subtests which typically make up different proportions of the subjects' scores at different ages or at different levels of competence.

It should be noted, however, that the methods of factor analysis do not reveal directly the nature of the factors isolated, nor do they provide complete assurance that these factors retain the same psychological meaning when the correlation matrix from which they have been derived is based on a different group of subjects. Identification and designation of the factors are based chiefly on a logical examination of their respective weights in the different items or subtests. For example, after a process known as *rotation of axes*, which is designed to organize the data into their simplest possible form,⁹ if it is found that within an original battery of twenty tests, Nos. 3, 8, 11, 13, and 18 have very high loadings for Factor I, and Nos. 1, 6, 7, 14, 15, and 20 have small loadings for that factor while Nos. 2, 4, 5, 9, 10, 12, 16, 17, and 19 have zero or negative loadings, one would naturally be led to ask in what respects the tests comprising the first series most strongly resemble each other in kind and

⁹ Guilford (1936) has presented an account of the methods of deriving factor loadings through the rotation of axes, which requires nothing more than an acquaintance with the principles of elementary algebra for its comprehension. Since his book was written, however, much additional work has been done in the field of factorial methods, particularly with respect to the concepts and methods underlying the principles of "simple structure." The later work, however, does not involve any fundamental change in the procedures as outlined by Guilford. For a more recent and detailed account of the methods and their mathematical bases the reader is referred to Thurstone's 1947 monograph.

are most unlike those in the last series. It might be found that those in the first series all require some sort of arithmetical reasoning. Nothing of this kind is demanded by those in the last series, which make their chief demand upon an understanding of word meanings. The tests in the middle group require some understanding of numbers and their relations but for the most part of a simpler and more formalized character, such as is needed for rapid and accurate computation or other stereotyped processes into which numbers enter. Under these conditions a tentative designation of Factor I as "arithmetical reasoning" seems reasonably justified. Other factors are identified and named in the same way or may be left unnamed, pending further investigation if the evidence seems insufficient to permit identification at the time.

When factors have been identified within one matrix of tests and the items or parts most heavily loaded or "saturated" with each have been indicated, the possibility of developing tests that will be "purer" measures of these underlying mental factors and which therefore can be interpreted with fewer hazards immediately comes to mind. A new battery is then made up including those items or subtests which showed the highest saturation with the factor selected for development, together with others thought promising for the purpose, and a factorial process is carried out with the revised series. This process is repeated until a point is reached at which all the variance except that which may reasonably be imputed to chance can be accounted for by a single factor. The name originally proposed for this factor may be retained or if, as sometimes happens, the later work should indicate that some other title gives a more exact meaning for the test, this may be substituted.

Thurstone is chiefly responsible, not only for developing the method of building tests presumed to measure these simpler and more unitary mental abilities, but also for carrying out the procedure to the point of devising actual tests. His original series (1938), which was intended for use with older children and adults, included tests of the following "primary" mental abilities: spatial (S), perceptual (P), numerical (N), verbal relations (V), memory (M), words, that is, single unrelated words (W), induction (I), reasoning (R), deduction (D). Three additional abilities were indicated by the analysis, but the evidence was not sufficient to warrant naming them. A more recent series has been devised for use with children of ages five and six (Thurstone and Thurstone, 1946). Five primary abilities were isolated at these ages, viz.: verbal-meaning (V), perceptual-speed (P), quantitative (Q), motor (M), and space (S).

All these, Thurstone has been careful to point out, are first approximations only. It is hoped that they will aid in reaching a valid description of the nature of mental organization which is at the same time reduced

to a small enough number of terms to bring it within the range of ordinary comprehension. Moreover, if the most basic factors underlying human abilities can be objectively measured, even though many of the finer nuances remain unnoted, a long stride toward a firmly rooted system of educational and vocational guidance and industrial classification will have been taken.

The methods of factor analysis can also be of considerable aid in answering two questions that have been raised before. First, how shall we define the limits set by the test content; in other words, just how far and in what directions does the universe of abilities measured by a particular test extend? A careful factorial analysis of the test items will go far toward answering this question. Second, how can we best determine the limits of the universe of subjects to whom a given test is applicable and for whom the obtained results will have similar meaning? If it is found that the factor pattern undergoes a greater change than can reasonably be accounted for by chance when the tests are given to groups of different composition from that of the one originally used in determining the factor loadings, it becomes decidedly open to question whether the second group should be regarded as belonging to the original universe. The use of the same normative standards for both is thus likely to be very misleading, since a given score earned by the members of one group has not the same psychological meaning as the same score earned by those belonging to the other group. The fact that the average score made by certain groups run consistently lower or higher than those earned by another group is no indication in itself that the test in question is not equally valid for both, for one group may be truly superior to the other. But a reliable difference in the factor patterns of the two is a different matter, since it suggests a basic difference in the pattern of mental organization underlying the test scores and accordingly a qualitative difference in the kind of abilities measured by the test which may or may not be accompanied by a difference in degree or level, since the two sets of scores cannot be regarded as comparable.

CONCLUDING REMARKS

The brief discussion in this chapter merely touches upon some of the more obvious questions relating to the examination of tests and testing methods in order to throw more light upon their intrinsic accuracy and their suitability for the purpose at hand. Enough may, however, have been said to show how great are the gaps in the needed information regarding most of the tests in common use and how few of the assumptions commonly made concerning these tests have been adequately based

upon objective facts. The rapid multiplication of new tests which has marked the history of testing during the past thirty or more years appears to be slowing up. It is to be hoped that the slower rate of production may be accompanied by greater care in the standardization of those which do appear, and that the more promising of the tests now in use may be subjected to further analysis which will make it possible to use them more appropriately and with better understanding of their results.

In an earlier section of this chapter it was noted that the most basic feature of any test is its validity for some specified purpose; that mere accuracy of measurement is of little account unless the measurement itself serves some practical or theoretical purpose. This statement can now be somewhat modified. The most important requirement of any test is that its psychological meaning shall have been soundly established; that we know just what it signifies. In the light of such knowledge we can apply it to appropriate situations and persons; in the absence of valid information of this kind, blunders and false conclusions are well-nigh inevitable.

Testing the Tests. II. The Divergence of Facts from Hypotheses

THE NULL HYPOTHESIS

One of the most basic principles of formal logic may be stated as follows: It is impossible to prove a universal negative, for the discovery of a single positive instance destroys the universality of the rule.¹ This principle is recognized in a number of practical situations. In debating, for example, the rule that the burden of proof rests with those who take the affirmative side of the question is well established. Those on the negative side need do nothing more than refute the arguments of their opponents; it is not necessary for them to demonstrate that the opposite state of affairs exists.

For the application of this principle to statistical methods we are chiefly indebted to R. A. Fisher. Inasmuch as it is possible to adduce evidence in support of a theory, whereas evidence against it cannot be regarded as proof that the theory is incorrect but only that its correctness has not been proved, Fisher suggested that scientific propositions should be formulated in negative terms and that experiments should then be set up to see whether or not these negative statements could be refuted by positive evidence. Such statements are known as "null hypotheses." The use of this method in the study of educational and psychological problems has become increasingly popular during recent years. For example, one might be interested in ascertaining whether or not the scores on a particular

¹ One might, of course, express the principle in the opposite way and say that it is equally impossible to prove a universal positive. This, however, would be the case only if it were possible to ascertain with complete assurance that a given instance is *not* positive or is devoid of positive features. To do this is theoretically impossible for there is always a chance that new and improved methods of study may reveal facts not now known. Apart from metaphysical considerations, we may say that the presence of a fact or condition is something that can be objectively determined, but its apparent absence may be an artifact resulting from imperfect means of detecting its presence. **Bacteria** existed before the microscope, but their existence was not known.

test given to engaged couples shortly before marriage are of any service in predicting happiness in marriage. The null hypothesis would then be phrased somewhat as follows: Scores on the ——— test obtained before marriage are *not* significantly related to happiness after marriage. "Significantly," as used here, means "to an extent unlikely to occur by chance." But since likelihood is not an all-or-none affair but has varying degrees, it becomes necessary to set an arbitrary limit to indicate how improbable an event must be to satisfy the dictates of caution. Just what level is selected will depend on the nature of the problem and the prudence of the investigator. Following the practice established by Fisher, it is customary to express probability in terms of chances in one hundred, that is, in percentages. In the example just given, if it should be found that the relationship between test score and subsequent marital happiness, as reported by the subjects in a single sample of cases selected to be representative of some larger universe, was high enough so that there were only two chances in one hundred that other samples from the same universe would either fail to show any relationship at all or perhaps show some trend in the opposition direction, the results obtained by the sample taken would be described as "significant at the 2 per cent level of confidence."²

The null hypothesis may then be *refuted* with varying degrees of assurance although its *correctness* cannot be established. All that can be said in favor of the hypothesis is that certain attempts at refutation have

² It is unfortunate that many persons who have adopted this method of expressing statistical results have lost sight of the continuous nature of the probability integral. Having decided to accept a certain level of probability as "significant," they make this a hard and fast boundary line with no regard for other factors which might affect the results. If they have decided on a 5 per cent level of probability as their criterion of "significance," then all findings which meet that criterion are for them as the laws of the Medes and Persians, while results which fall short of the magic figure, be it only by a fractional per cent (say 5.25 chances in a hundred instead of the 5.00 which was chosen as the maximum), are thrown aside as "not significant." Here again, as in the case of the word "reliable" previously discussed, the choice of a form of expression having a well-established meaning in everyday speech seems unfortunate.

There are a number of statistical procedures for which the probabilities of obtaining results of various magnitudes are laborious to compute. For most of these, tables are available from which these probabilities may be read directly. As a rule, however, economy of space has led to the reduction of these tables to the values necessary for certain levels of probability; in most cases only the .20, .10, .05, .02, and .01 levels being indicated. In the shorter tables only the 5, 2, and 1 per cent points may be shown.

The selection of these particular points for inclusion in the published tables was, of course, based merely upon arithmetical convenience. But to many who use the tables, these points seem to have taken on a kind of sacrosanct character in comparison to which the levels falling beyond and between them are of slight consequence. This has too often led to the reporting of statistical results merely as "significant" or "not significant" in place of giving the actual figures for what they may be worth.

failed. It is, of course, true that in the practical situation, if repeated attempts by all known methods continue to show negative results, presumption in favor of the hypothesis is strengthened. But this is not the same thing as the positive evidence which may be accumulated against the hypothesis and for which the probabilities of error may be determined. Proof (within its limits) is a positive thing that takes us farther along the road to understanding; absence of proof leaves us pretty much where we were before. The statistical formulation of the null hypothesis has done a good deal to clarify thinking in this area and has made for better design of experiments in many scientific fields. Its rapid gain in popularity is evidence of its practical usefulness as well as of its scientific advantages.

DEGREES OF FREEDOM

If a stick twelve inches long is to be divided into two parts, it is not possible to decide what the length of one part shall be without considering that of the other part. If the first part is eight inches in length, the second automatically becomes four inches. Although two pieces have been made, only one choice of lengths was possible. In statistical terminology, only a single *degree of freedom* existed. If the stick had been cut into three pieces, two free choices could have been made, but the length of the third piece would have been determined by the lengths chosen for the other two. Two degrees of freedom would have been involved.

The number of degrees of freedom in any problem depends upon the conditions that have been imposed. If the only condition is that the totals of the parts shall conform to some fixed value (twelve inches in the example just given), the number of degrees of freedom will be one less than the number of parts. But if additional restrictions are set by the nature of the problem, the number of degrees of freedom is correspondingly reduced. Suppose, for example, that a series of test items is being examined with a view to selecting those which differentiate reliably between mechanics and traveling salesmen. The null hypothesis requires an assumption that there is no difference between the two groups in their ability to deal with the items under consideration. In order to see whether or not the hypothesis is refuted, the results obtained for an experimental sample consisting, let us say, of 60 mechanics and 40 salesmen may be compared, item by item. For Item 1 the results might be those shown below:

	<i>Pass</i>	<i>Fail</i>
Salesmen	30	10
Mechanics	25	35

If the null hypothesis is exactly satisfied, there will be only one degree of freedom in such a table, for the percentages of success and failure must then be the same in both groups. Since there are 100 responses in all, of which 55 are "passes," it becomes possible to convert the figures in this table into those which would have been found if the hypothesis had been fulfilled. In that case, exactly $55/100$ or 55 per cent of each group would have passed the item and 45 per cent would have failed with it. The results as they would have been had the hypothesis been satisfied are shown below:

	<i>Pass</i>	<i>Fail</i>
Salesmen	22	18
Mechanics	33	27

It will be noted that when the total number of successes is known and the condition of equiproportionality is imposed, the determination of one of the four figures making up the table automatically fixes the value of the other three. Thus only one degree of freedom exists. Had there been three occupational classes instead of two (say mechanics, salesmen, and clerical workers), two degrees of freedom would have obtained, since it would be necessary to establish values for two of the occupational groups before that of the third becomes automatically fixed. And if, instead of classifying the responses merely as "passed" or "failed," a five-way system of values had been employed (such as "excellent," "good," "average," "poor," or "very poor"), then there would have been four degrees of freedom in the columns (just as when dividing a stick into five parts there is free choice of length for four of the divisions but the length of the fifth is fixed by that of the other four) and two degrees of freedom for the rows (three occupational classes) or $4 \times 2 = 8$ degrees of freedom for the entire table.

CHI-SQUARE (χ^2) AND THE NULL HYPOTHESIS

Returning now to the problem set in the preceding section, we note at once that the actual proportions of success and failure on the item under consideration by the salesmen and the mechanics, respectively, do not conform to those which would have been found had the null hypothesis been exactly upheld. The question that remains to be answered is: How much confidence may we attach to the apparent finding that salesmen do better than mechanics on Item 1? In other words, would this item be a useful one to include in a test designed to classify individuals with respect to their relative aptitude for the two fields of work?

The first step is to find the differences between the observed and the theoretical frequencies for each cell in the table. Since there is but one degree of freedom, these differences will be the same for each cell; in this case, 8. Each divergence is then squared and divided by the theoretical value for the cell. The sum of these proportions gives us a measure of the total divergence of the observed results from the postulated condition (no difference between the two occupational groups) in their ability to succeed with the item in question. The figures then become

$$\frac{8^2}{22} + \frac{8^2}{18} + \frac{8^2}{33} + \frac{8^2}{27} = 10.78.$$

The result is known as Chi-square (χ^2).³ As can be seen from the example given, χ^2 is the sum of the ratios of the squares of the divergences in each cell from the theoretical values set by the null hypothesis to the theoretical number of cases in the cell. We need to know its sampling distribution, that is, what the chances are of obtaining a χ^2 value of any given size if no actual differences exists between the variables in question within the universe from which the samples are drawn. For this we have reference to a table originally worked out by Karl Pearson in 1900 to which certain corrections were made by "Student," and which was later put into more convenient form by R. A. Fisher. An abbreviated form of this table is given on page 237; for more exact data the reader should consult the more complete presentation in Fisher's *Statistical methods for research workers* or the *Statistical tables* by Fisher and Yates. Tables of the χ^2 distribution are also to be found in many of the textbooks on statistical methods (Lindquist, 1940; Guilford, 1936; Peters and Van Voorhis, 1940; and others).

Reference to the *Statistical tables* shows that with one degree of freedom a χ^2 as large as 6.635 would occur by reason of chance sampling only once in a hundred trials and that one as large as 10.827 would occur only once in a thousand times. In the case just considered, $\chi^2 = 10.78$, which barely falls short of the latter figure. The null hypothesis is thus refuted with a very high degree of confidence and the inclusion of the item in a test designed for the selection of traveling salesmen seems well justified.

It should be noted that χ^2 is not a measure of the extent of the

³ The formula for Chi-square is

$$\chi^2 = \sum \frac{(f_o - f_t)^2}{f_t}$$

where f_o is the observed frequency in any category; f_t is the corresponding theoretical frequency; and Σ means "the sum of."

TABLE 6
 χ^2 VALUES FOR DIFFERENT DEGREES OF FREEDOM*
 LEVELS OF CONFIDENCE

d.f.	50%	20%	10%	5%	2%	1%	d.f.	50%	20%	10%	5%	2%	1%
1	.455	1.642	2.706	3.841	5.412	6.635	16	15.338	20.465	23.542	26.296	29.633	32.000
2	1.386	3.219	4.605	5.991	7.824	9.210	17	16.338	21.615	24.769	27.587	30.995	33.409
3	2.366	4.642	6.251	7.815	9.837	11.341	18	17.338	22.760	25.989	28.869	32.346	34.805
4	3.357	5.989	7.779	9.488	11.668	13.277	19	18.338	23.900	27.204	30.144	33.687	36.191
5	4.351	7.289	9.236	11.070	13.388	15.086	20	19.337	25.038	28.412	31.410	35.020	37.566
6	5.348	8.558	10.645	12.592	15.033	16.812	21	20.337	26.171	29.615	32.671	36.343	38.932
7	6.346	9.803	12.017	14.067	16.622	18.475	22	21.337	27.301	30.813	33.924	37.659	40.289
8	7.344	11.030	13.362	15.507	18.168	20.090	23	22.337	28.429	32.007	35.172	38.968	41.638
9	8.343	12.242	14.684	16.919	19.679	21.666	24	23.337	29.553	33.196	36.415	40.270	42.980
10	9.342	13.442	15.987	18.307	21.161	23.209	25	24.337	30.675	34.382	37.652	41.566	44.314
11	10.341	14.631	17.275	19.675	22.618	24.725	26	25.336	31.795	35.563	38.885	42.856	45.642
12	11.340	15.812	18.549	21.026	24.054	26.217	27	26.336	32.912	36.741	40.113	44.140	46.963
13	12.340	16.985	19.812	22.362	25.472	27.688	28	27.336	34.027	37.916	41.337	45.419	48.278
14	13.339	18.151	21.064	23.685	26.873	29.141	29	28.336	35.139	39.087	42.557	46.693	49.588
15	14.339	19.311	22.307	24.996	28.259	30.578	30	29.336	36.250	40.256	43.773	47.962	50.892

* Table 6 is abridged from Table III of Fisher: *Statistical methods for research workers*, Oliver & Boyd Ltd., Edinburgh, by permission of the author and publishers. A more complete table is to be found in *Statistical tables* by R. A. Fisher and F. Yates, published by Oliver & Boyd, Edinburgh.

difference between two groups but only of the degree of assurance with which one may state that *some* difference exists. For this reason, the method has fewer applications to mental testing than to other types of experimental work, but it is frequently helpful in the selection of test items and in the testing of hypotheses with respect to group differences or to bias in test content. Moreover, since the method is based upon the proportion of observed to expected frequencies, a sufficient number of cases is needed to render these proportions fairly stable. Lindquist (1940) suggests that the minimum number of cases in the sample should not be less than 50 and that no cell should have a theoretical frequency smaller than 10. These figures, of course, are only approximate.

As a measure of the extent to which the facts obtained for a single sample of cases diverge from those to be expected on the basis of the null hypothesis, or any other specific hypothesis that may be set up, the use of Chi-square is not confined to a 2 by 2 celled table but may be used with any number of rows and columns. It is important in such cases to enter the probability table with the correct number of degrees of freedom which, it will be remembered, is commonly the product of the number of rows less one multiplied by the number of columns less one. Thus in a table with six rows and four columns there would be fifteen degrees of freedom— $(6-1)(4-1)$.

Although Chi-square does not provide a measure of the extent of relationship between two variables,⁴ the coefficient of contingency derived from it provides such a measure. The contingency coefficient (*C*) is used to ascertain the degree of relationship between two sets of attributes which are expressed in categorical rather than quantitative terms. For example, common observation would lead us to expect some relationship between the religious preferences of mothers and daughters, but observation alone does not enable us to state the strength of the association. And since religious preferences do not form a quantitative series but only a list of categories arranged in arbitrary order, neither the product moment nor the correlation ratio method⁵ of determining the relationship is applicable.

The coefficient of mean square contingency, more briefly designated

⁴ Paradoxical as it may sound, a measure of difference is at the same time an indication of relationship. If a reliable difference is found between the scores made by the sexes on some test, then there is a correlation between these scores and sex. If measures increase with age, there will be a difference between the means of successive ages and likewise a correlation between the measures and the age of the subjects. The form of expression differs but the facts are the same.

⁵ See Chapter 17.

as the contingency coefficient, is obtained by means of the formula given below:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

where N is the number of cases. Unless the number within each category is large enough to lend reasonable dependability to the data, the method should not be used.

Values of the contingency coefficient do not correspond exactly to those of the Pearsonian r , although they approximate them more nearly as the number of categories is increased. As obtained, the sign is always positive, but an inspection of the scatter diagram will usually reveal whether the indicated relationship takes the same or an opposite direction. Thus, in the example given, if it should be found that the daughters of Methodist mothers show a greater than chance likelihood of preferring the Methodist Church to other denominations and similarly for the other categories, the relationship would be positive, but if there were a dependable tendency for Methodism in the mothers to be associated with avoidance of that denomination in the daughters to an extent that could not reasonably be accounted for on the basis of chance, the coefficient (though still bearing the positive sign) would be interpreted as having a negative significance. The magnitude of the contingency coefficient ranges from zero (indicating no relationship) to an upper limit equal to the square root of the reciprocal of the number of categories subtracted from 1.00. If there are six categories, the highest possible value of the contingency coefficient would be

$$\sqrt{1 - 1/6} = \sqrt{.8333} = .913.$$

DIFFERENCES BETWEEN MEANS OF CONTINUOUS VARIATES

In mental testing, particularly when dealing with results obtained by the use of tests which have already been standardized and for which no further modification in content is feasible, one is more often interested in comparing the average performances of different groups on the test as a whole than in considering merely the relative proportions of those who "pass" or "fail" the test according to some arbitrarily established criterion. Our question then becomes: With how much confidence can we say, on the basis of obtained samples drawn from each, that Group A is superior to Group B in respect to performance on this test? Again we phrase our problem in terms of the null hypothesis (no difference between the groups) and ascertain with how much confidence we are warranted in refuting it—how certain we may be that other samples

drawn from the same universes would show a difference in the same direction (though not necessarily of the same magnitude) as that found in the two samples we have actually compared.

The question hinges upon these points: (1) the size of the samples, (2) the extent of the differences between the two means, (3) the extent of correlation between scores in the two samples, and (4) the spread of

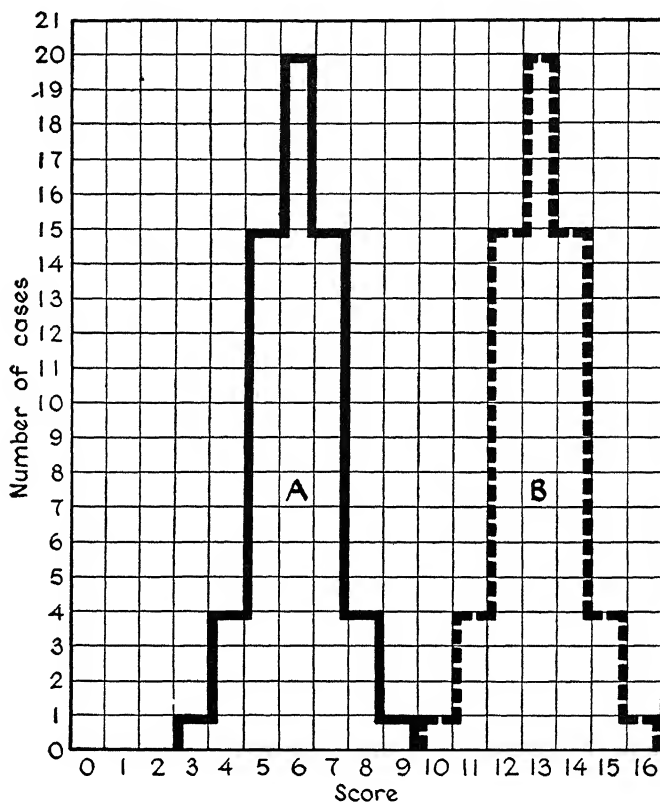


FIG. 13. NONOVERLAPPING SCORES OF TWO RELATIVELY HOMOGENEOUS GROUPS.

the scores in each group about its own mean. If this spread is very small in each case, then there may be no overlapping at all between groups, even though the means are separated only by what seems like a rather moderate amount. The highest case in the lower group would still fall below the lowest case in the upper group. (See Figure 13.) From their scores alone, one could safely place all the cases in each of the two samples in their proper classes. But if the spread of scores had been greater with the means remaining the same, as shown in Figure 14, then

some of the cases in Group B would earn scores that fall within the range covered by Group A, and some of those in Group A will do no better than a few of the best in Group B. It would thus be unsafe to try to classify the subjects on the basis of their scores alone, even though the means of the two classes are as widely separated as those in the example first cited. Because of the overlapping between the groups, we can be less certain that other samples would show a mean difference similar to the one found in this case.

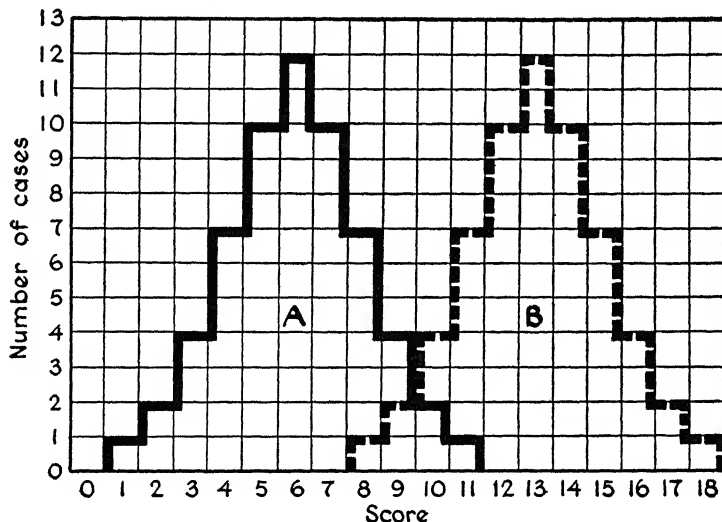


FIG. 14. OVERLAPPING SCORES OF TWO MORE HETEROGENEOUS GROUPS WHOSE MEAN SCORES DIFFER BY THE SAME AMOUNT AS THAT SHOWN IN FIGURE 13.

In order to ascertain the amount of confidence with which we may assume that a difference actually found between the means of two samples is a valid indicator of real superiority of the one group over the other, that is, the likelihood that a similar difference would be found in other samples, we then need to know, first, how much variation in the means of successive samples we may expect to find if each is drawn from the same universe; and, second, how high is the correlation between the scores made by the two groups that are compared. Inasmuch as in the majority of instances it will not be feasible to determine the variability of the means of a large number of samples by actual experiment, a statistical estimate is substituted that is based on the variability of the individual scores about the mean of the single sample which was actually obtained. As was previously noted, if individuals differ greatly, one from another, the means of successive samples are likely to shift

more from sample to sample than would be the case if all the subjects earned scores that were more nearly alike.

The best estimate that can be made of the probable distribution of the means of a series of samples from a given population on the basis of the distribution of scores on a single sample drawn from that population is given by the following formula:

$$\sigma_M = \frac{S.D._{dist.}}{\sqrt{N-1}}$$

where $S.D._{dist.}$ is the standard deviation of the distribution of scores in the sample and N is the number of cases. Since the number of differences between successive scores is one less than the number of scores (see previous section on degrees of freedom), modern statisticians prefer the formula given above to that likely to be found in the older textbooks where \sqrt{N} is used as the denominator of the fraction. Of course if N is sufficiently large, it makes little difference which formula is used.

From a knowledge of the chance distribution of the means of samples from a given universe, we are enabled to pass directly to a second problem of major importance in the field of mental testing. Suppose that we have two samples, each from a physically different universe which may or may not be similar with respect to some measured characteristic, say scores on a particular test. The two distributions of scores are not identical and their means are not equal. With how much confidence may we reject the hypothesis that the universes from which these samples were drawn do not differ in respect to scores on the test in question? In other words, what are the chances of finding similar differences between further samples drawn at random from each of the two universes separately?

To answer this question we first need to know the standard errors of the means of the two samples. We also need to know whether or not the scores in the two distributions are correlated. Since correlation means that a part of their variances are common to both, any difference between them must be attributable to the part which remains. This lessens the probability of large differences, and makes any divergence in the means more significant than it would be if the variance in each were wholly independent of the other. The formula for determining the standard error of a difference between two means thus becomes

$$\sigma_{diff.} = \sqrt{\sigma_{M1}^2 + \sigma_{M2}^2 - 2r_{12}\sigma_{M1}\sigma_{M2}}$$

where σ_{M1} and σ_{M2} are the standard errors of the two means as given by the previous formula, and r_{12} is the correlation between the scores. If the variables are uncorrelated ($r = 0$), then the final term, of course, drops out and is disregarded.

In the older textbooks on statistical method, the method recommended for determining the probability of exceeding a given divergence was based upon the assumption that the variance of a sample affords the best available estimate of the variance of the universe from which it is drawn. While this is approximately true for large samples, it does not hold for smaller ones. Actually, the variance of a sample is likely to be *smaller* than the variance of its universe, and the difference between the two increases rapidly as the number of cases in the sample diminishes. For samples of thirty or more cases, the differences have become small enough so that they may be safely disregarded (see Table 4, page 195), but when the sample is small, the improbability ("significance") of a given divergence will be spuriously exaggerated by the use of the older method, which consisted of dividing the difference between the means of the two groups to be compared by the standard error of the difference as determined by the formula just given. The result, known as the "critical ratio" (C.R.), is interpreted by reference to the probability table (Table 4). Since it indicates the number of standard deviations by which the obtained difference exceeds the average of a distribution of the differences in a theoretical series of paired samples drawn in a manner similar to that used for the pair on which the computation was based, the probability (p) can be read directly from the column of deviations (x values) corresponding to the size of the C.R. For large samples, this procedure is sufficiently accurate. But when the number of cases is small, not exceeding 30, the discrepancy between the true and the obtained estimates of probability becomes large enough to have a serious effect upon the conclusions likely to be drawn.

An English statistician who modestly signs himself merely as "Student" was the first to call attention to the effect of this divergence between the respective variabilities of sample and universe upon the sampling distribution of the means of small samples, and its consequent effect upon comparisons of the means of two or more groups. The computation needed for correction would be laborious were it not for the fact that R. A. Fisher has devised a substitute for the procedure just described (the critical ratio method) to which he has given the name of the t statistic. Although it is based on the difference formula previously given, it introduces a correction which allows for the variability factor. The formula for t is given below:

$$t = \frac{M_1 - M_2}{\sqrt{\left(\frac{\Sigma d_1^2 + \Sigma d_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

where M_1 and M_2 are the means of the two samples;

Σd_1^2 and Σd_2^2 are the sums of the squares of the deviations of the individual measures from their respective means; and N_1 and N_2 are the numbers of cases in each sample.

The sampling distribution of t is given in Table 7. The number of degrees of freedom will be equal to the number of cases in the two samples combined less 2, that is, the sum of the d.f.'s for each. If there are 8 cases in one group and 12 in the other, the total will have $(8 - 1) + (12 - 1) = 18$ d.f.

The table should be entered with the appropriate number of degrees of freedom and read across until a point is reached at which the obtained value of t corresponds most nearly to that given in the table. The level of confidence warranted by this value is shown at the top of the column. Interpolations may be made when desired.

TABLE 7
LEVELS OF CONFIDENCE FOR VALUES OF t AT SUCCESSIVE DEGREES OF FREEDOM*
LEVELS OF CONFIDENCE

d.f.	50%	20%	10%	5%	2%	1%
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.051	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
12	0.695	1.356	1.782	2.179	2.681	3.055
14	0.692	1.345	1.761	2.145	2.624	2.977
16	0.690	1.337	1.746	2.120	2.583	2.921
18	0.688	1.330	1.734	2.101	2.552	2.878
20	0.687	1.325	1.725	2.086	2.528	2.845
22	0.686	1.321	1.717	2.074	2.508	2.819
24	0.685	1.318	1.711	2.064	2.492	2.797
26	0.684	1.315	1.706	2.056	2.479	2.779
28	0.683	1.313	1.701	2.048	2.467	2.763
30	0.683	1.310	1.697	2.042	2.457	2.750
∞	0.675	1.282	1.645	1.960	2.326	2.576

* Table 7 is abridged from Table IV of Fisher: *Statistical methods for research workers*, Oliver & Boyd Ltd., Edinburgh, by permission of the author and publishers. Attention is called to the more complete table given therein and to the larger collection to be found in *Statistical tables* by Fisher and Yates published by Oliver & Boyd, Edinburgh.

Sometimes the interest centers about a comparison of the variability of scores made by two groups rather than in the differences in their means. For example, as was shown in Chapter 11, if intelligence quotients are to have the same meaning at all ages, their variability at the different ages must be equal. Even a relatively small difference in the spread of the scores from one age to another will mean large differences in the proportions at the extremes.

One of the very live issues in psychological testing at the present time has to do with the question of the comparative variability of the sexes in mental ability. Even though the average performance of two groups is equal, if one varies more from its own average performance than does the other, the proportion of cases with very high and very low ranks will be greater in the more variable group. It is unquestionably true, for example, that the proportion of famous men greatly exceeds the proportion of famous women, and that a similar though less marked difference between the sexes can also be noted among those who have achieved better than average success in the learned professions or in the upper ranks of business and industry. In his study of gifted children in California, Terman found that among elementary school children the number with IQ's of 140 or over was in the ratio of 116 boys for every 100 girls. Among the high school children of corresponding mental level the ratio was 183 boys to 100 girls (Terman, *et al.*, 1925). Macmeeken (1939), who gave the 1937 Stanford-Binet to all the children in Scotland whose tenth birthdays fell on one of four specified dates, obtained a mean IQ for boys of 100.5, S.D. 15.9. For the girls the corresponding figures were 99.7, S.D. 15.2. Small as these differences are, the resulting sex ratio among those with IQ's of 140 or higher would be 158 boys to 100 girls. Whether these differences can best be explained on the basis of test content or whether a more basic difference exists in the variability of the sexes cannot be said with complete assurance.

Whenever interest centers about extreme cases rather than on the average of groups—and these instances are many⁶—the question of variability may become even more important than that of average level of performance. The measure of variability commonly used in such studies is the standard deviation, and again a difference is made between large and small samples in determining the likelihood that the direction of an obtained difference between two measured samples of a given universe will be maintained in succeeding trials by the same procedure. The

⁶ For example, in the study of delinquents and criminals, sex perverts, persons showing exceptional talent or deficiencies along some particular line, the relative proportions of the mentally defective or those with exceptional intellectual gifts in certain specified groups, etc.

probability that an obtained difference between the standard deviations of two samples expressed in terms of the standard error of that difference may be determined in the usual way from the probability table if the sample is large (say 50 or more cases). It is necessary to find the most probable value of the standard deviation of a series of samples drawn from the same universe, that is, the standard error of the standard deviation. The formula commonly used for samples of 50 or more is

$$\sigma_{S.D.} = \frac{S.D.}{\sqrt{2N}}$$

where N is the number of cases. These values are then substituted in the formula given on page 242 for finding the standard error of the difference between two means, and the number of times by which the difference between the standard deviations of the two samples to be compared exceeds the standard error of that difference is ascertained. Reference to the probability table will then show the level of confidence with which the null hypothesis is refuted.

As in the case of the means, however, sampling differences are not normally distributed. The departure from perfect normality becomes less and less as the size of the sample increases. While it may safely be ignored in samples of, say, 50 or larger, rather serious misinterpretations may arise from disregarding this factor when dealing with small samples. In such cases the *variance ratio* to which Snedecor gave the name of the F statistic⁷ should be used. The formula is based on the ratio between the estimated variance of the standard deviations of successive samples drawn from each of the two universes:

$$F = \frac{\frac{\sum d_1^2}{N_1 - 1}}{\frac{\sum d_2^2}{N_2 - 1}}$$

where d_1 and d_2 are the differences between the individual scores and the mean in each of the two samples;

N_1 and N_2 are the corresponding numbers of cases; and

Σ means "the sum of."

The sampling distribution of F is shown in the table on page 248. The ratio is always taken in such a manner that the larger variance is in the numerator. Its value is therefore never less than 1.00.

It should be noted that in an occasional case where t is found to reach a level of significance great enough to render the null hypothesis

⁷ In honor of R. A. Fisher, who first worked out the formula upon which its sampling distribution is based.

highly improbable, the discrepancy may be due to a difference in variability rather than to a difference in the means. If there is reason to suspect that this is the case, the F test may be applied as a check. It is possible that both may differ.

The distribution of the statistics of small samples is sufficiently different from that of the larger samples, upon which the methods presented in many of the standard textbooks are based, to necessitate careful examination of much of the data reported in the psychological and educational literature. Even in current publications, one need not search far to find critical ratios derived by the large-sample procedure and interpreted according to the normal probability curve when the number of cases does not exceed 20 and may even be as few as 10. The conclusions reached with respect to confidence levels are always exaggerated in these cases, and the use of the proper methods will often show them to be invalid.

A word of caution is needed, however, with respect to the use of small samples when human subjects are concerned. The statistics are sound, but the psychology may be questionable. To many people, the demonstration of what can be accomplished through careful planning of the experimental setup when very small samples of plants and animals are used as subjects has been convincing evidence that the larger samples formerly thought necessary can be dispensed with. They lose sight of certain facts that should be obvious: (1) that human beings vary along more coordinates than do plants or animals, especially in respect to mental life and behavior; (2) that the conditions which give rise to these variations are less well known and but little subject to experimental control; and (3) their behavior is to a greater extent self-initiated and self-directed. All these factors make the universe of human behavior less homogeneous; its subjects differ more, one from another. To be representative of such a universe, a sample must cover a wider area; it must be made up of more parts. When small samples are all that can be obtained, or in preliminary work when an experimenter wishes to check his progress, the use of these methods is certainly called for, but their greater precision for such purposes should not blind us to the fact that a small sample of a large and complicated universe is still likely to be an inadequate representation of that universe. No statistical device, however sound it may be, can make up for lack of basic data.

THE PRACTICAL USE OF TESTS OF "SIGNIFICANCE" IN MENTAL MEASUREMENT

In relatively few instances is the interest in a particular comparison limited to the particular groups which are compared. Much more often

TABLE 8

VALUES OF F AT THE 1 PER CENT (BOLD FACE TYPE) AND 5 PER CENT (LIGHT FACE TYPE) LEVELS OF CONFIDENCE FOR DIFFERING DEGREES OF FREEDOM

d.f. ₂		Degrees of Freedom for Larger Variance (d.f. ₁)													
		1	2	3	4	5	6	7	8	9	10	20	50	100	∞
Degrees of Freedom for Smaller Variance	*1	161 4052	200 4999	216 5403	225 5625	230 5764	234 5859	237 5928	239 5981	241 6022	242 6056	248 6208	252 6302	253 6334	254 6366
	2	18.5 98.5	19.0 99.0	19.2 99.2	19.3 99.3	19.3 99.3	19.3 99.3	19.4 99.3	19.4 99.4	19.4 99.4	19.4 99.4	19.4 99.5	19.5 99.5	19.5 99.5	19.5 99.5
	3	10.1 34.1	9.6 30.8	9.3 29.5	9.1 28.7	9.0 28.2	8.9 27.9	8.9 27.7	8.8 27.5	8.8 27.3	8.8 27.2	8.7 26.7	8.6 26.4	8.6 26.2	8.5 26.1
	4	7.7 21.2	6.9 18.0	6.6 16.7	6.4 16.0	6.3 15.5	6.2 15.2	6.1 15.0	6.0 14.8	6.0 14.7	6.0 14.5	5.8 14.0	5.7 13.7	5.7 13.6	5.6 13.5
	5	6.6 16.3	5.8 13.3	5.4 12.1	5.2 11.4	5.1 11.0	5.0 10.7	4.9 10.5	4.8 10.3	4.8 10.2	4.7 10.1	4.6 9.6	4.4 9.2	4.4 9.1	4.4 9.0
	6	6.0 13.7	5.1 10.9	4.8 9.8	4.5 9.2	4.4 8.8	4.3 8.5	4.2 8.3	4.2 8.1	4.1 8.0	4.1 7.9	3.8 7.4	3.8 7.1	3.7 7.0	3.7 6.9
	7	5.6 12.3	4.7 9.6	4.4 8.5	4.1 7.9	4.0 7.5	3.9 7.2	3.8 7.0	3.7 6.8	3.7 6.7	3.6 6.6	3.4 6.2	3.3 5.9	3.3 5.8	3.2 5.7
	8	5.3 11.3	4.5 8.7	4.1 7.6	3.8 7.0	3.7 6.6	3.6 6.4	3.5 6.2	3.4 6.0	3.4 5.9	3.3 5.8	3.2 5.4	3.0 5.1	3.0 5.0	2.9 4.9
	9	5.1 10.6	4.3 8.0	3.9 7.0	3.6 6.4	3.5 6.1	3.4 5.8	3.3 5.6	3.2 5.5	3.2 5.4	3.1 5.3	2.9 4.8	2.8 4.5	2.8 4.4	2.7 4.3
	10	5.0 10.0	4.1 7.6	3.7 6.6	3.5 6.0	3.3 5.6	3.2 5.4	3.1 5.2	3.1 5.1	3.0 5.0	3.0 4.9	2.8 4.4	2.6 4.1	2.6 4.0	2.5 3.9
	12	4.8 9.3	3.9 6.9	3.5 6.0	3.3 5.4	3.1 5.1	3.0 4.8	2.9 4.7	2.9 4.5	2.8 4.4	2.8 4.3	2.5 3.9	2.4 3.6	2.4 3.5	2.3 3.4
	14	4.6 8.9	3.7 6.5	3.3 5.6	3.1 5.0	3.0 4.7	2.9 4.5	2.8 4.3	2.7 4.1	2.7 4.0	2.6 3.9	2.4 3.5	2.2 3.2	2.2 3.1	2.1 3.0
	16	4.5 8.5	3.6 6.2	3.2 5.3	3.0 4.8	2.9 4.4	2.7 4.2	2.7 4.0	2.6 3.9	2.5 3.8	2.5 3.7	2.3 3.3	2.1 3.0	2.1 2.9	2.0 2.8
	18	4.4 8.3	3.6 6.0	3.2 5.1	2.9 4.6	2.8 4.3	2.7 4.0	2.6 3.9	2.5 3.7	2.5 3.6	2.4 3.5	2.2 3.1	2.0 2.8	2.0 2.7	1.9 2.6
	20	4.4 8.1	3.5 5.9	3.1 4.9	2.9 4.4	2.7 4.1	2.6 3.9	2.5 3.8	2.5 3.6	2.4 3.5	2.4 3.4	2.1 2.9	2.0 2.6	1.9 2.5	1.8 2.4
	25	4.2 7.8	3.4 5.6	3.0 4.7	2.8 4.2	2.6 3.9	2.5 3.6	2.4 3.5	2.3 3.3	2.3 3.2	2.2 3.1	2.0 2.7	1.8 2.5	1.8 2.3	1.7 2.2
30	4.2 7.6	3.3 5.4	2.9 4.5	2.7 4.0	2.5 3.7	2.4 3.5	2.3 3.3	2.3 3.2	2.2 3.1	2.2 3.0	1.9 2.6	1.8 2.2	1.7 2.1	1.6 2.0	
50	4.0 7.2	3.2 5.1	2.8 4.2	2.6 3.7	2.4 3.4	2.3 3.2	2.2 3.0	2.1 2.9	2.1 2.8	2.0 2.7	1.8 2.3	1.6 1.9	1.5 1.8	1.4 1.7	
100	3.9 6.9	3.1 4.8	2.7 4.0	2.5 3.5	2.3 3.2	2.2 3.0	2.1 2.8	2.0 2.7	2.0 2.6	1.9 2.5	1.7 2.1	1.5 1.7	1.4 1.6	1.3 1.4	
∞	3.8 6.6	3.0 4.6	2.6 3.8	2.4 3.3	2.2 3.0	2.1 2.8	2.0 2.6	1.9 2.5	1.9 2.4	1.8 2.3	1.6 1.9	1.4 1.5	1.2 1.4	1.0 1.0	

we wish to draw conclusions with respect to larger universes of which the groups in question are regarded as samples. Because a given sample is rarely an exact duplicate in miniature of the universe from which it is drawn and also because different samples drawn from the same universe will usually differ from each other to a greater or less degree, it becomes a matter of prime importance to make the best possible estimate of the probability that similar results would be obtained in other investigations of the same kind. We need to know not only what results are obtained in a particular investigation but also with how much confidence we may assume that these results would be duplicated in other studies made in like manner.

Another important aspect of the determination of confidence levels is seen in those instances where the conditions of experiment require that certain numerical relationships be maintained. An example already mentioned is that, in scales standardized by the quotient method, these quotients must show equal variability from age to age if their meaning is to be uniform at all ages. In actual practice, however, some variability of the standard deviations is to be expected as a result of differences in sampling. It thus becomes necessary to inquire in each case how likely it is that an obtained difference merely represents a chance fluctuation of sampling within two homologous universes or whether the difference is so great that the null hypothesis cannot reasonably be upheld. A determination of the variance ratio (the F test) will provide an answer to the question. Had this test been generally applied, some of the irregularities in the results obtained from a number of the tests now in use could have been foreseen and corrected.

The question of sex differences in performance on a given test poses another problem. Should the differences found be looked upon as indicators of actual differences in the relative ability of the sexes or do they result from bias in content of the tests? As far as single tasks or test items are concerned, results may be taken at their face value, but when a complex test made up of many separate items or tasks is involved, the problem becomes more difficult. Two courses are open in such cases,

FOOTNOTE TO TABLE 8

* The upper figure in each pair shows the value required for the 5% level; the lower figure indicates that required for the 1% level. To enter the table, find the column corresponding to the number of degrees of freedom for the larger of the two variances, and the row corresponding to the number for the smaller variance. The intersection of the two will show the values of F needed for each of the two levels of confidence indicated. Interpolation between the values given may be made when necessary.

This table is a somewhat abbreviated form of the complete table presented by G. W. Snedecor in *Statistical methods*. Fourth Edition. Ames: Iowa State College Press, 1946. It is reproduced here by permission of the author and the publisher.

each with its own set of assumptions. In some instances the test constructor may be willing to trust his own judgment as to the choice of tasks that are equally "fair" for both sexes. If it then turns out that one sex makes a reliably better showing than the other, it is assumed that a corresponding difference between the sexes exists in respect to the universe of which the test is presumed to be a representative sample—that there is a sex difference in the "trait." In other instances an artificial leveling of the performances of the two sexes may be achieved by purposely selecting items on which males and females do equally well or by taking an equal number of those which favor each of the sexes. When the first procedure is followed, the burden of proof rests on the hypothesis that males and females have truly had equal opportunity and incentive to acquire the skills demanded by those items on which one or the other is found to excel; in the second case the assumption of actual sex equality must be tested by some method which does not involve an artificial canceling of whatever differences may exist. It is impossible to lay down rules for testing these assumptions since crucial experiments will depend upon the nature of the universe to be studied, but if the problem is clearly recognized, some way of resolving the dilemma will usually present itself. The danger lies in drawing conclusions without a clear realization of the premises on which these conclusions depend.

Not only sex differences but differences between groups separated on bases other than sex are subject to the same hazards of interpretation when general traits or attributes that manifest themselves in diverse ways are to be appraised by means of tests that include a variety of different items. Whether the experimenter is aware of it or not, the use of such tests always involves certain hypothetical assumptions which may or may not be warranted. The testing of hypotheses in terms of the probability of exceeding a given divergence in a sample from its theoretical value has become a well-recognized aspect of modern statistical method, but such testing is too often confined to the final results of an investigation and fails to take account of the more basic hypotheses implicit in the method chosen for arriving at those results. Unless the foundation is sound, no structure can be depended on to stand.

Correlated Measures

BASES OF CORRELATION

Correlated measures are those which show concomitant variation, either in the same or in the opposite direction from each other. If high scores in one of the correlated measures tend to be associated with high scores in the other, the correlation is positive; if high scores in the first are accompanied by low scores in the second, the correlation is negative. When neither tendency can be observed, the correlation is zero. In most cases a negative correlation can be changed to a positive one by merely altering the form in which scores are expressed. For example, it is usually found that the greater the degree of skill that has been attained, the shorter will be the time required to perform an act. The relationship between *time required* and proficiency of performance is therefore a negative one. But the same relationship may be expressed in positive terms by simply altering the form of expression and changing the time scores to their reciprocals, in which case we should say that a positive relationship exists between proficiency and *speed* of performance. Likewise a negative relationship (though probably not a rectilinear one) has usually been found between intelligence in women and sexual promiscuity, but this may likewise be changed to the positive form by a simple reversal of the scores in the first variable so that backwardness instead of superiority is matched with the measures of promiscuity. We then say that a positive relationship has been found between the degree of *mental retardation* and the likelihood or the extent of sexual promiscuity in women. Because most people find it easier to think of a positive relationship than of a negative one, it is usually better to arrange the scores and to express the results in positive terms unless this makes for too great awkwardness in the form of expression necessitated.

Although there is often a strong temptation to infer causation from correlation, to do so is usually hazardous. As a matter of fact, most scientists are very hesitant about referring to cause and effect, and some are unwilling to do so at all, preferring to speak only of antecedents and

consequents. The word "cause," however, is a brief and convenient one which need occasion no misunderstanding if it is recognized as not implying either a circumstance which by itself constitutes an adequate explanation of an event or condition, or one that inevitably leads to a given result. By a *cause* we mean only something that occurred earlier in time from which later events might be predicted with more than chance probability. Only in this sense are we warranted in saying that causative factors may be inferred from correlation. If the correlated variables are of such a nature that one necessarily precedes the other, it may be convenient to say that the first was a cause (not the sole cause unless the correlation is perfect, and not the original cause in any event) of the second. This means only that from a knowledge of the first, one could predict, with a degree of assurance indicated by the magnitude of the correlation coefficient, that the second event would follow. If no time relationship is involved, we may think of the two as being concomitant results of a common underlying factor, or, more precisely, of a series of factors.

One of the dangers of inferring cause from correlation is the likelihood that curiosity will thereby be satisfied and investigation be brought to an end. When it was (erroneously) believed that the majority of juvenile delinquents were mentally deficient, little further study of other possible factors leading to delinquent behavior was made for a number of years, since lack of intelligence was deemed a sufficient explanation. Even today there are many people who are satisfied with finding that many of the groups who consistently make low scores on intelligence tests have come from relatively poor environments. Lack of environmental stimulation or opportunity is accepted as an adequate explanation or *cause* for the low test standing, and only when this explanation is challenged by those who take a different position with respect to this question is further research thought necessary. On the other hand there are persons who adopt much the same attitude with respect to the color of the skin. Negroes do poorly on intelligence tests because they are Negroes; further explanation is considered unnecessary. Intelligence test scores of blood relatives are positively correlated, and the closer the degree of relationship the higher is the correlation. Thus it is assumed that intellectual variations are *caused* by inherited factors of a biological nature, and there are many who are willing that the question should end there.

Like most statistical problems, correlation resolves itself into a question of probabilities; the chances of error in drawing conclusions with respect to unknown facts on the basis of those which are known. If this point is kept in mind, it becomes clear that the popular idea of

cause and effect, in the sense that one of the variables operates as an agent in producing or inducing variations in the second, is seen to be both unnecessary and confusing. Correlation merely affords a guide to prediction and to the amount of confidence that may be placed in the predictions made. "Cause" and "effect" are terms which should be used sparingly if at all; when employed it should be only in the limited and special sense here indicated.

DEPENDENT AND INDEPENDENT VARIABLES

In the use of correlation for predictive purposes, the most probable score in one variable is estimated on the basis of the score made on the other variable and the degree of correlation between the two measures, all scores being expressed in standard units. The measure on which the scores are known is called the *independent* variable; that on which the scores are to be predicted, the *dependent* variable. In some cases it is a matter of choice which variable is to be regarded as the independent and which as the dependent variable, a fact which has occasionally led to confused thinking and even to ridiculous conclusions. For example, in 1921 Cyril Burt presented a series of correlations with their regression equations from which he drew the conclusion that more than half of the variance of school children on a modified form of the Binet tests could be attributed to school attainment. He arrived at this conclusion by calculating the partial regressions of his measurements on Binet mental age as the dependent variable, which yielded a regression weight of .54 as the contribution of schooling to mental age. In criticizing this conclusion, Holzinger and Freeman (1925) demonstrated that when the same data were used and the same line of reasoning was followed, it could be shown that the variance in chronological age, taken as the dependent variable, is more than half "attributable" to school attainment (regression weight = .51). They remarked that such a state of affairs is truly alarming, for if children of superior educational attainments are to grow old at so much more rapid a rate than the generality, the advantages of schooling may well be questioned!

This is a neat illustration of the hazards of imputing a causal relationship, as the term is popularly used, to the results obtained by a correlational analysis. If regarded merely as means of predicting one series of measures from another, both of the equations just mentioned are valid and equally so. It is certainly possible to improve one's estimate of what a child's Binet mental age is likely to be if one knows his educational age, which was the measure of school attainment used by Burt. It is equally possible to arrive at an estimate of his chronological

age on the basis of the same data, nor is it unreasonable to find that the improvement over a chance guess which is made possible by such information is about equally great in both instances.

A special case in which recognition of the difference between dependent and independent variables becomes important is that in which the subjects are *selected* on the basis of high or low scores in one of the variables to be compared. More will be said about this in a later chapter, but it is well to note at this point that because of the method used in selection, the errors of measurement in the variable used in making the selection (the independent variable) will not be distributed at random but will be greater in the direction toward which the selection was made. If cases making higher scores than the generality are chosen, not only the true scores but errors taking a positive direction will operate in producing these scores. Errors in the opposite direction occur, of course, but since their effect is to lower the obtained scores below their true value, there is less chance that such cases will be chosen, even though their true scores would have led to their inclusion in the group. Thus whenever a biased selection of this kind is made, as is frequently done when contrasted groups are to be compared, it must not be forgotten that the scores earned on the independent variable will on the average be too high or too low accordingly as the selection was made on the basis of superior or inferior performance on the measure in question. But since errors are presumed to be uncorrelated,¹ the scores in the dependent variable will ordinarily not be affected. The errors will be as likely to take one direction as the other, and if the group is large they may be expected to cancel each other, in which case the obtained mean will not differ greatly from the true mean. A direct comparison of the means of the two groups or of other measures into which the means enter thus becomes invalid. Many instances in which ignorance of this principle has led to very misleading conclusions are to be found in the experimental literature. Outstanding examples are some of the studies on the use of the so-called *accomplishment ratio* or *accomplishment quotient*, which will be described in Chapter 22, and on the effect of various factors upon the intelligence quotient.

SOME CORRELATES OF INTELLIGENCE

The principle that correlation rather than compensation is the rule with respect to most if not all mental traits is nowhere better

¹ As has been noted previously, the assumption of uncorrelated errors is not always justified. When correlation between errors exists, some effect of the selective method will appear in the dependent as well as in the independent variable, but this effect will ordinarily be much less marked.

exemplified than in the realm of intelligence, at least insofar as intelligence is indicated by mental test scores. The common belief that exceptionally bright children are likely to become stupid as they grow older, or at least to become impractical and eccentric adults, as well as the corresponding belief that most eminent men were dunces in school, has been thoroughly disproved by objective study of the facts. In the report by Terman and Oden (1947) on the adult achievements of approximately 1500 cases who had been selected in childhood on the basis of their exceptionally high standing on mental tests, it is shown that not only was the early high standing maintained² as far as measures of their intellectual ability are concerned, but that they were also superior to the generality of people in health and in personal-social characteristics. High childhood intelligence is thus not counterbalanced by later stupidity, loss of health, or social peculiarities.

Correlation rather than compensation has been shown to hold good at the opposite end of the scale of intelligence. Stupidity in childhood is unlikely to be compensated for by mental brilliance later on. The degree of self-correlation between measurements of mental ability at different ages varies with a number of factors, but with the exception of those taken in infancy, it is always positive.

With intelligence³ kept as the independent variable, it has been found that most other measurable characteristics show at least a low positive correlation with the test scores. Even physical measurements share in this tendency, though the correlations are usually not high. Tests of personal-social characteristics follow the same rule, with high scores on the intelligence tests likely to be associated with "desirable" scores on the tests of personal-social traits and vice versa. These relationships, however, are not as certainly established as many others, and their interpretation is open to some question since superior ability may enable the subjects to perceive the purpose of these tests and to select the socially desirable answers on an intellectual basis rather than in terms of the trait which the test was intended to measure. To the extent that this is done, the correlations would be spuriously raised. The relatively low self-correlations of most of the tests of this kind must also be taken into account. Without correction for attenuation, correlations with other measures are necessarily low because of the high experimental error in the scores on the dependent variable. But if the first type of error is

² As noted in the preceding section, some falling off in later test standing was to be expected on the later tests because of the elimination of bias in the direction taken by the errors of measurement on the first occasion.

³ For the sake of brevity we have sometimes used the term "intelligence" as if it were synonymous with "mental-test score," which is, of course, the variable actually meant.

also present, the use of the attenuation formula may overcorrect to such an extent that the result will be more misleading than would have been the case if the original correlation had been accepted at its face value.

Practically all measures of educational achievement show high correlation with intelligence-test scores, and particularly so when the measure of intelligence is a group test of the usual linguistic kind. As a matter of fact, the relation between scores on many group intelligence tests and those earned on a good battery of school achievement tests is so high that Kelley (1927, page 209) came to the conclusion that "most of the distinctions drawn between intelligence and achievement are spurious." It should be noted, however, that the measure of intelligence used by Kelley was the National Intelligence Test (1920), which was modeled after the Army Alpha test used in World War I. The subtests of which the N.I.T. is made up include a number of the same kind as are used in others described as "educational tests," such as arithmetical reasoning, the completion of sentences by filling in omitted words, and a test of word meaning in which the child is required to say whether the members of a pair of words have the same or the opposite meaning. All except one of the subtests in each form make some demand upon reading ability.

It is, however, quite to be expected that with an educational system such as prevails throughout most of the United States, a genuinely high relationship should be found between intelligence, as usually conceived and defined, and educational achievement. Broadly speaking, we may say that intelligence makes school achievement possible when suitable opportunity and incentives for the latter are provided, and when differences in special talent which give rise to differential performances in the various subject fields are roughly smoothed out by taking a composite of the scores in all the subjects, rather than just a single one, as the measure of achievement.

Something of the same reasoning may explain the oft-demonstrated fact that most measures of environmental conditions are also related to scores on intelligence tests. Amount of income, number of books in the home, various measures of social standing, class of neighborhood in which the home is located, size of home and quality of furnishings, and a host of other external factors have been found to be correlated with the intellectual level both of parents and of their children as well. It is not unreasonable to assume that the more intelligent members of the population, other factors being equal, are better able to procure these benefits for themselves and for their children than are the less intelligent, who are not so quick to recognize or avail themselves of opportunities, or to foresee the consequences of their acts.

That bright children tend to be superior to the generality in health and in physical size may possibly be due to some general underlying factor which, for want of a better name, we may designate as "the general quality of the organism," but it may also be the result of better physical care on the part of their parents. It will be remembered that on the average the parents of bright children are also above average in mental ability and are more likely to have the financial means which enable them to provide healthful living conditions for their children. There are, of course, many exceptions to this rule, and it would be enlightening to compare the correlations between health and intelligence or those between intelligence and physical measurements for children from homes where the expected agreement with respect to mental ability and socioeconomic status⁴ is found, with the corresponding relationships for those cases where the opposite conditions exist (bright children coming from poor homes, backward children from good homes). In order to avoid spurious correlation due to the "halo" effect of unconsciously taking scores in one variable into account when rating another,

* The question is: Can the usual positive but low relationship between intelligence-test scores and physical status be accounted for in terms of environmental factors or must we postulate some more basic type of relationship? The proposed test consists in separating the available subjects into two groups of which the first (Group A) would include those subjects whose intelligence corresponds at least roughly to their home background (bright children coming from good homes, backward ones from poor homes), and the second (Group B) would include those for whom the opposite conditions exist. If the relationship between intelligence and physique, as found for the children, can be adequately accounted for in terms of the quality of their home care, one would expect this correlation to be positive for Group A and negative for Group B.

A better test, if cases are available, would be a comparison between pairs of siblings of like sex and tested at corresponding ages. In this case the procedure would consist in making a comparison of the means in one of the two variables for groups divided on the basis of the other variables. For example, if the relationship between IQ and height is to be tested, the pairs should be separated in such a way that the child with the higher IQ is always placed in group A; the one whose IQ is lower, in Group B. Since it is reasonable to assume that, on the whole, the home care given to children in the same family is not radically different, if it is found that the children in Group B not only have a lower average IQ than those in Group A (as they must because of the way the groups were divided), but also tend to be shorter, some relation between intelligence and physique over and above that due to environment is indicated. A check on this relationship may be had by reversing the basis of selection. The taller members of each pair may be placed in Group A; the shorter ones in Group B. The grouping will not be the same as in the first test because of the low correlation between the two variables. If the mean IQ of the children in Group A is higher than that of those in Group B, further evidence for the above relationship is thereby afforded. The double check is desirable because it is unlikely that either difference, taken by itself, will be large enough to meet the usual requirements for statistical dependability. But if the double check is made for each of several independent groups, the consistency with which the *direction* of the results is maintained, even though the differences are small, may provide sufficient evidence.

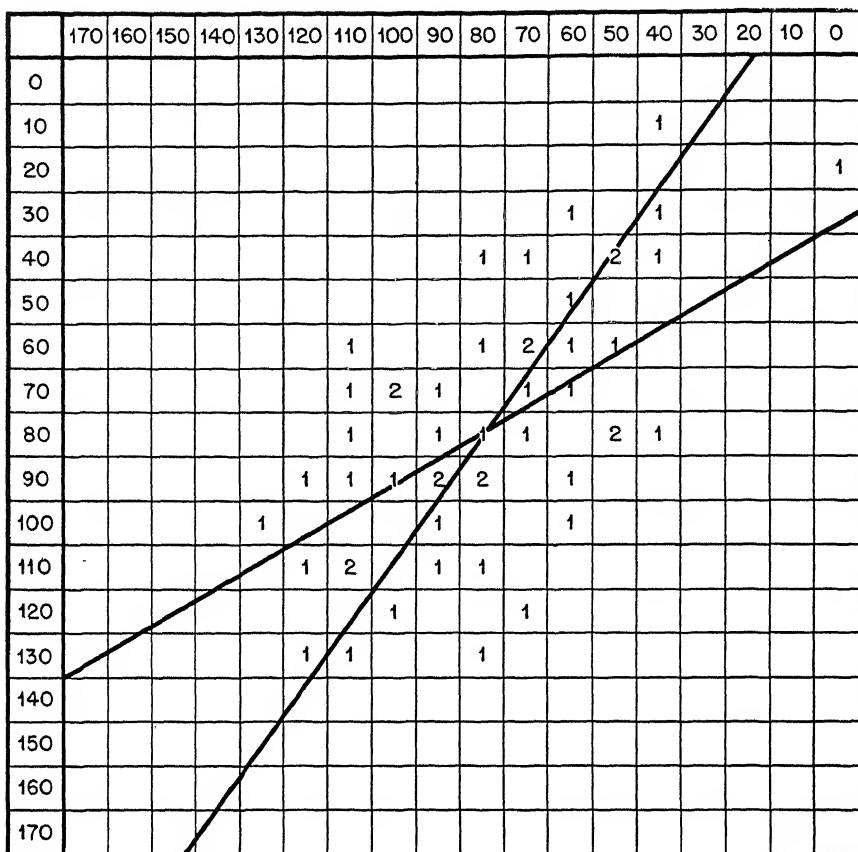


FIG. 15. CORRELATION BETWEEN SCORES ON TWO FORMS OF AN INTEREST TEST
WHEN $r = +.60$.

the classification of homes as "good" or "poor" should be done by someone who does not know either the purpose of the study or the IQ's of the children.

FURTHER METHODS OF DETERMINING RELATIONSHIPS

Thus far the only method of ascertaining the extent of relationship between two variables that has been considered is the product-moment method of correlation, known as the Pearsonian r , for which the formula was given on page 163. It would be beyond the scope of this book to enter into the various methods of correlational analysis in detail; we shall

simply mention briefly a few of the procedures used when the nature of the data is such as to render the product-moment method unsuitable.

The valid use of the product-moment method requires that the line of regression drawn through the means of the successive rows and also that drawn through the means of the columns in a correlation surface

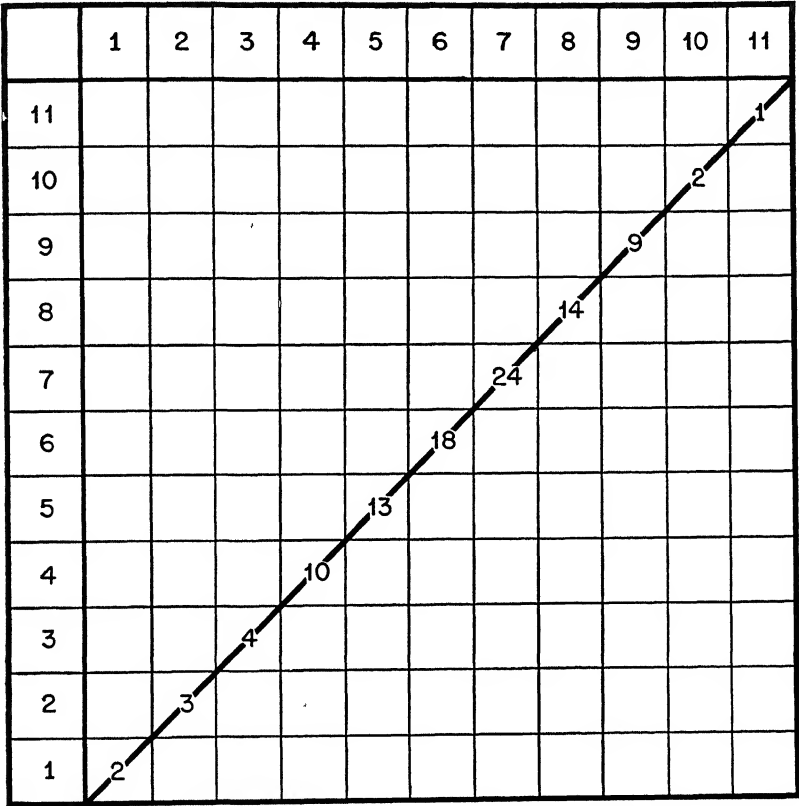


FIG. 16. DISTRIBUTION OF SCORES ON A SCATTER DIAGRAM WHEN $r = +1.00$.

such as that shown in Figure 15 shall be straight, within the limits of their standard errors. Particularly is it important that they show no consistent tendency to curvature. The distribution of scores found within a single line or column of a correlation surface (also known as a “scatter diagram”) is called an *array*. It is customary to refer to the columns as the x arrays and to the measure of which the scores are placed in the columns as the x variable. The measure of which the scores are entered in the rows is called the y variable. The requirement that the means of the arrays in each variable shall form an approximately straight line is

known as the rule of *rectilinear regression*,⁵ and if this condition is not fulfilled, the product-moment method will not show the true relationship between the variables.⁶ In such cases the *correlation ratio*, which is a more generally applicable method, should be used.

If correlation is rectilinear and perfect ($r = 1.00$), all the scores in each variable will fall within the cells through which the regression lines pass (see Figure 16), and the standard deviations of the *means* of

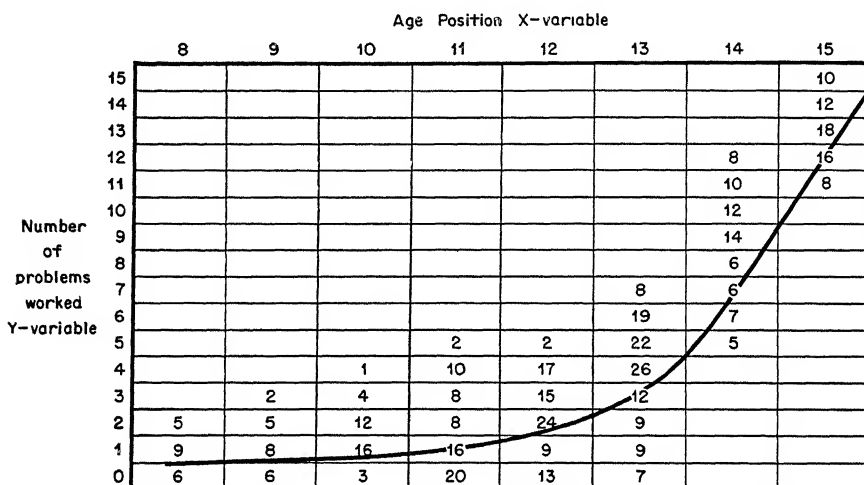


FIG. 17. CURVILINEAR REGRESSION SHOWN BY SCORES ON A TEST WHICH IS TOO DIFFICULT FOR THE MAJORITY OF CHILDREN BEFORE THE AGE OF TEN OR ELEVEN YEARS.

the x (or the y) arrays will be equal to that of the *scores* on the x (or the y) variable. If the correlation is rectilinear but less than 1.00, it will be equal to the ratio of these values, that is, the ratio of the standard deviations of the means of the x arrays to that of the x scores or the corresponding figure as calculated for the y 's. If the regression is truly rectilinear, these ratios will not differ from each other or from Pearson's r by an amount greater than might be expected by chance. But if the difference is greater than can reasonably be accounted for by chance, then the use of a single figure to express the relationship is misleading, for the deviations of the scores from the line of best fit, in other words, the error of estimating one score from another, will be greater in one direction than in the other. (See Figure 17.) Although the correlation

⁵ Sometimes called *linear regression*.

⁶ Other requirements of the product-moment method as strictly interpreted, such as *homoscedasticity* (equal standard deviations of the arrays) and *homocliticity* (symmetry in the form of the arrays), will not be discussed here since they are of relatively minor importance unless the departure from the required conditions is marked (Kelley, 1923, p. 172).

ratios, as these values are called, do not permit an algebraic estimation of the probable score on the dependent variable as does the regression equation based on Pearson's r , but merely indicate how close is the maximum relationship between the two variables, a graphic solution of the problem may be had by plotting a smooth curve through the means of the successive arrays of the independent variable and reading across to find what the corresponding scores on the dependent variable is most likely to be. By reversing the process, so that the measure formerly regarded as the independent variable becomes the dependent one and vice versa, the estimate can be made in the other direction.

The correlation ratio is a more general method of expressing relationship than Pearson's r , since it is applicable to all quantitatively expressed data regardless of the form of the regression line. However, it is rarely used when the product-moment method is equally permissible, since r is easier to compute and has properties that the ratio method lacks. The symbol for the correlation ratio is the lower-case Greek eta (η), with subscripts to indicate which is the independent variable and which the dependent. The letter designating the latter is written first. Thus η_{yx} indicates the regression of the y scores on the x variable. The magnitude of the correlation ratio is a measure of the goodness of fit, that is, an indication of the extent to which the y scores conform to the line drawn through the means of the x arrays. If most of the scores fall on or very close to that line, it is possible to make a very close prediction of what a given y score will be from a knowledge of its corresponding x score, and η will therefore be high, perhaps coming very close to its maximum value of 1.00. But if the y scores scatter widely on both sides of the x line, then a given value of x will provide only a very undependable approximation to the corresponding value of y and the correlation ratio will be low, perhaps not far from zero.⁷

The formulas for the two correlation ratios are given below:

$$\eta_{xy} = \frac{S.D.M_x}{S.D._x}$$

and

$$\eta_{yx} = \frac{S.D.M_y}{S.D._y}$$

where $S.D.M_x$ and $S.D.M_y$ are the standard deviations of the *means* of the x and y arrays, respectively, and $S.D._x$ and $S.D._y$ are the standard deviations of the individual scores in the two distributions.

⁷ As computed, the correlation ratio always has a positive sign. Inspection of the figures or of the diagram showing the regression line will indicate whether high scores in one variable go with high scores in the other or the reverse.

A very clear and straightforward account of the steps in the computation of the correlation ratios and of the corrections for grouping is given by Garrett (1947). Discussions of the method and its use will also be found in most of the texts on statistical method.

Since η is computed from the actual line of regression while r is calculated from the straight line that most nearly approximates it, the former can never be lower and will usually be higher than the latter. The assumption of rectilinear regression, however, has the great advantage that the slope of the line is determined by all the data, and if the total number of cases is not large, this lends a stability to the results which is wanting when an attempt is made to follow all the vagaries in a regression line of which some of the apparent shifts are in all likelihood due to chance. For example, if the cases are so few and the grouping is so fine that there is only one entry in each array of the diagram, then scattering of scores becomes impossible and, no matter what the true relationship may be, the correlation ratio will have a value of 1.00, which in such a case is meaningless. In general, therefore, it is unwise to use this method unless there are enough cases to permit a grouping of the data into a sufficiently large number of classes to show the trend of the relationship with enough entries in each class to lend some degree of stability to the location of the means. Kelley (1923) has described a method of correcting the correlation ratio for grouping that is too fine (leading to too few cases in each array and therefore to a correlation ratio that is too high) and also for grouping that is too coarse, which makes for lack of discrimination within the arrays and hence to a spurious lowering of the ratio. Of these, the first is much the more important.

When there is reason to suspect the hypothesis that the lines of regression do not depart significantly from rectilinearity (as is required when the product-moment method is to be used), an analysis of the variance of each measure into two components, that due to rectilinear regression and that due to departure from rectilinearity, should be made. (See Chapter 18.) If the F test shows the second component to be significantly greater than zero, the correlation ratio should ordinarily be used in preference to r .⁸ In deciding which method to use when the issue is doubtful, particularly when the number of cases is not large enough to make it possible to organize the data in such a way that the correlation ratios can be satisfactorily freed from errors of grouping, and more especially when the problem involved is of such a nature that an overestimation of the relationship would lead to more serious conse-

⁸ Blakeman's test of rectilinearity, which is given in most of the older textbooks on statistical methods, has been shown by R. A. Fisher to be inadequate since it does not take account of the number of arrays.

quences than would an underestimate, the product-moment method may be chosen even though the odds against perfect rectilinearity of the regression line are as much as 20 to 1 (critical ratio 1.64). But if conservatism takes the opposite tack, making an underestimation of the relationship the thing to be most avoided, then the correlation ratio, which will always yield the higher figure except in those rare instances in which the lines of regression for the measured sample chance to be *absolutely* straight, should be used if there is any strong suggestion of nonrectilinearity, unless, of course, the number of cases is very small. Here, as elsewhere, statistical procedures should always be dictated by common sense and an appreciation of the requirements of the problem at hand.

Spearman's rank-difference method provides a convenient way of determining the approximate degree of correlation between two variables when the number of cases is not large. Because the labor involved in computing it increases very rapidly with the number of cases involved, it is but little used with population samples exceeding twenty-five cases. The method was originally devised for determining the relationship between judgments of the rank order of persons or products in terms of some specified criterion. It may, however, be used with quantitative measures by first arranging the cases in the order of their scores on each of the two variables separately, numbering them in order of rank, and substituting the rank numbers for the original scores. If two or more cases have the same score, the ranks are averaged. If, for example, the two persons occupying fifth and sixth position have equal scores, each is assigned a rank midway between 5 and 6, which is 5.5.

The next step is to find the differences between the ranks on the two variables for each case separately, disregarding signs. These differences are first squared, then the sum of the squares is multiplied by 6, and the product divided by an amount equal to $n(n^2 - 1)$ where n is the number of cases. To obtain the rank-difference correlation, this quotient is then subtracted from 1.00. The complete formula thus becomes

$$\rho = 1 - \frac{6\sum d^2}{N(N^2 - 1)}$$

where $\sum d^2$ is the sum of the squares of the differences between ranks and N is the number of cases.

The chief advantage of the rank-difference method lies in the fact that when the number of cases is small, the correlation can be computed more quickly by its use than by any other equally valid way. In the course of an investigation it frequently becomes desirable to have a basis for tentative conclusions respecting probable relationships as the work

progresses. Often such information obtained at succeeding intervals will suggest new lines of approach and will enable the investigator to avoid following will-o'-the-wisps too long.

The probable error of ρ (ρ), which is the symbol of a correlation computed by the rank-difference method, is about 5 per cent greater than that of r for the same data would be when quantitative scores are converted into ranks, provided that no ranks have been averaged. The effect of averaging ranks varies with the number averaged, the true value of the relationship as indicated by r , and the position in the series of the ranks that are averaged. It can be shown, however, that if all the scores in one series are averaged, thus giving them equal rank, while all those in the other series retain their separate ordinal positions, the value of ρ becomes .50, which is, of course, meaningless. It follows that, on the whole, if the true value of the correlation is higher than .50, the most likely effect of averaging ranks will be to lower the value of ρ (bringing it nearer to .50), while if the true value is below .50, averaging ranks is more likely to bring about a spurious increase in the value of ρ .

Even apart from chance fluctuations or from differences arising from averaging ranks, correlations obtained by the rank-difference method do not correspond exactly to product-moment correlations. The difference, however, is not large. Tables for transmuting ρ coefficients into their most probable values of r will be found in most textbooks on statistical methods. For practical purposes it is sufficient to remember that r will average about .02 higher than ρ when the latter correlations run between (approximately) .35 and .65, and about .01 higher for values of ρ above and below this range.

The method known as *biserial r* is resorted to when one of the variables is expressed in the usual continuous form while the other is restricted into two classes. The valid use of this method requires that the latter condition is merely the result of inadequate information or crude measurement. That the dichotomous variable is actually continuous and approximately normal in form of distribution must be not only a reasonable but the most likely assumption if the method of biserial r is to be used. For example, a school supervisor may wish to know how well scores on a certain reading test given at the beginning of the school year will predict promotion into the next grade at the close of the year. Although the second variable is divided into only two classes—promotion or non-promotion—the achievement for which it stands is by no means a two-way affair, but each term covers a wide range. Among those who are promoted we shall have some of outstanding merit and others who just barely manage to squeeze through. Among those who fail of promotion there will be some who did almost as well as the poorest of those who

passed and others whose achievement was so poor that promotion was out of the question. A similar but less obvious case in which the use of biserial r appears justified is in the item analysis of tests when either an outside criterion or the total score on the remainder of the test is used as a basis for evaluating the individual items. It might seem at first that when items are scored as "right" or "wrong" two discrete categories and not a continuous series are indicated. More careful consideration will lead to the conclusion that the situation is actually very similar to that of the previous example. The group whose members answer the item correctly will include a range extending all the way from those who could have answered much harder questions of the same class to those whose success was chiefly due to fortunate guessing. In like manner those marked "wrong" will cover many degrees of ignorance. But since the formula for biserial r^9 involves one of the properties (z) of the normal probability integral, the use of the formula when the dichotomous variable cannot reasonably be looked upon as normal in form is not permissible. If normality of the constricted variable cannot be assumed, a direct comparison of the mean scores of the two groups on the continuous variable is perhaps the soundest method to use. Biserial r_p , which, like the correlation ratio, makes no assumption of normality, is laborious to compute and is not often used.

Particularly in exploratory work, it may be desirable to make a comparison of extreme cases as a means of determining the extent of relationship of one variable to another. Not infrequently teachers, industrial foremen, or others who may be asked to make judgments respecting the persons whom they supervise, feel incompetent or lack the time to go through a formal rating system, but if asked merely to name the ten best and the ten poorest of these persons with no distinctions among them will do so readily enough. Peters and Van Voorhis (1940) have developed a method for handling material of this kind, which they call "biserial r from widespread classes." The same assumptions with respect to the true distribution of the judgments as in ordinary biserial r must be warranted. Only the tails of the distribution, however, are used; the middle section is omitted. If these tails are of unequal size with more "good" than

⁹ The formula for biserial r is

$$r_{\text{biser.}} = \frac{(M_2 - M_1)pq}{S.D.(z)}$$

where M_2 and M_1 are the mean scores on the distributed variable of the two groups respectively;

p and q are the proportions in each group (together making 100 per cent); $S.D.$ is the standard deviation of the entire distribution; and z is the ordinate of the normal curve at the point of dichotomy.

"poor" cases chosen or vice versa, or if it is not feasible to secure measures on the continuous variable with which the judgments are to be compared for the entire group or for a representative sample of it, the procedure becomes very laborious and will not be described here. But if the tails are of equal size¹⁰ and if it is feasible to give the test to all the subjects, the procedure is not difficult. The formula is

$$r_{wsp} = \frac{(M_2 - M_1)p}{2S.D.(z)}$$

where r_{wsp} is the correlation between scores on the test and judgments of excellence (or whatever criteria were used in choosing the tails);

M_1 and M_2 are the means of the test scores of the two selected groups; p is the proportion of the total group included in each tail;

$S.D.$ is the standard deviation of the test scores made by the entire group; and

z is the ordinate at the point where the tails are cut off.

This formula is particularly valuable for persons who are forced to ask the cooperation of busy persons or those whose acquaintance with the subjects to be judged is only moderately close when securing criteria for evaluating tests. The latter are likely to know the extreme cases but may be quite hazy in their impressions of the middle group; the former may know all the cases but lack the time to provide detailed information.

When both variables, instead of just one, are constricted into the dichotomous form,¹¹ tetrachoric correlation may be used to determine the relationship. Until the publication of the *Computing diagrams* by Chesire, Saffir, and Thurstone in 1933, this method was not often used because of the time required for computation, but since then it has become very popular, especially when factorial analysis is to be made of the items in a relatively long test. By the use of the *Diagrams*, all the intercorrelations of the items can be worked out in a small fraction of the time that would otherwise be required.

The methods here mentioned do not, by any means, include all the ways of studying relationships that are useful in mental measurement. No attempt has been made to do more than provide an extremely brief account of some of the procedures most commonly used and of the con-

¹⁰ There is no set requirement either as to the number of subjects or the proportion of the group to be included in the tails except that there must be a sufficient number to lend reasonable stability to the means. Ten is about the minimum figure; a larger number is preferable.

¹¹ The same assumptions must be made here as are required for the dichotomous variable in the biserial methods. The relatively high sampling error of the tetrachoric correlation must also be considered.

ditions requisite for their application to the construction and use of tests. We turn now to the question of confidence levels, to the likelihood of correlation between measurements of a sample drawn from a universe in which the attributes in question are uncorrelated, and to the determination of the limits within which it may be said (with a stated degree of assurance) that the correlation in the universe is likely to fall when that in the sample is known.¹²

Our first question may be phrased as follows: With how much confidence can we state that other samples drawn from the same universe will show a like correlation trend, i.e., one having the same sign but not necessarily of the same magnitude, as that observed in a single obtained sample? Are we warranted in rejecting the hypothesis that the correlation found in this sample was merely a chance departure from zero?

Since r is not normally distributed (see footnote 12), the t test, using the formula given below, may be applied. Reference to Table 13, page 212, of Lindquist's *Statistical analysis*, which shows the correlation needed to reach the 5 per cent and the 1 per cent levels of confidence for varying numbers of cases, makes this computation unnecessary since

¹² The formula for determining the standard error of r given in most of the older textbooks on statistical methods is

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}}.$$

This formula is incorrect since it assumes that the sampling errors of r are normally distributed, which is not the case. Even when $r =$ zero the variance of a small sample tends to be smaller than that of the universe from which it is drawn, and for values higher than zero the sampling distribution becomes markedly skewed even for samples of very large size. It will be recalled that the successive values of r do not represent equal steps on an arithmetically calibrated scale. An r of .80 does not indicate merely twice as close a relationship as one of .40; the difference is much greater than that. The increase from a correlation of .30 to a correlation of .50 is by no means comparable to that from .70 to .90. Fisher has shown that a measure derived from r which he calls the z function (this must not be confused with z as the height of the ordinate in the probability integral) overcomes both these difficulties since the sampling errors are normally distributed, or nearly so, and the units are equally spaced. Since the derivation of the z function is based upon logarithms and a separate value must be computed for each level of r , much labor is saved by the use of a table from which the transmuted values can be read directly. One of the most conveniently arranged of these tables is to be found in Lindquist's *Statistical analysis in education* (1940) of which an abbreviated form is presented in Table 9. It will be noted that when r is close to zero, z does not differ from it very much, but the difference is marked at the higher levels, since z increases by equal steps while r does not.

The standard error of z as given by R. A. Fisher is equal to

$$\sigma_z = \frac{1}{\sqrt{N - 3}}$$

where N is the number of cases in the sample.

TABLE 9

VALUES OF FISHER'S z FUNCTION FOR SUCCESSIVE VALUES OF r *

r	z	r	z	r	z	r	z	r	z
.00	.000	.21	.213	.42	.448	.63	.741	.84	1.221
.01	.010	.22	.224	.43	.460	.64	.758	.85	1.256
.02	.020	.23	.234	.44	.472	.65	.775	.86	1.293
.03	.030	.24	.245	.45	.485	.66	.793	.87	1.333
.04	.040	.25	.255	.46	.497	.67	.811	.88	1.376
.05	.050	.26	.266	.47	.510	.68	.830	.89	1.422
.06	.060	.27	.277	.48	.522	.69	.848	.90	1.472
.07	.070	.28	.288	.49	.536	.70	.867	.91	1.528
.08	.080	.29	.299	.50	.549	.71	.887	.92	1.589
.09	.090	.30	.310	.51	.563	.72	.908	.93	1.658
.10	.100	.31	.321	.52	.576	.73	.929	.94	1.738
.11	.110	.32	.332	.53	.590	.74	.950	.95	1.832
.12	.121	.33	.343	.54	.604	.75	.973	.955	1.886
.13	.131	.34	.354	.55	.618	.76	.996	.960	1.946
.14	.141	.35	.365	.56	.633	.77	1.020	.965	2.013
.15	.151	.36	.377	.57	.648	.78	1.045	.970	2.092
.16	.161	.37	.388	.58	.662	.79	1.071	.975	2.185
.17	.172	.38	.400	.59	.678	.80	1.100	.980	2.298
.18	.182	.39	.412	.60	.693	.81	1.127	.985	2.443
.19	.192	.40	.424	.61	.708	.82	1.157	.990	2.647
.20	.203	.41	.436	.62	.725	.83	1.188	.995	2.994

* Reproduced in abbreviated form from *Statistical analyses in educational research* by E. F. Lindquist, by permission of and by arrangement with the publishers, Houghton Mifflin Company.

For more exact work the reader is referred to the original table on p. 215 of the reference cited.

the figures may be read directly from the table.

$$t = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2}.$$

It needs but slight consideration to demonstrate that many of the "conclusions" based upon correlations of low or moderate size and small numbers of cases, which have been reported in the literature and for which standard errors have been computed by means of the older formula (see footnote 12), have but shaky foundation when the t test is applied. For example, an r of $+.40$ derived from a sample of 25 cases just reaches the 5 per cent level of confidence, but the erroneous assumption of normally distributed sampling errors made by the older formula would make it appear reliable at better than the 1 per cent level. Had there been but 20 cases, the correlation would still have been judged "sig-

nificant" at the 3 per cent level of confidence by the older method, but the t test would require an r of .444 to reach even the 5 per cent level. If there are but 10 cases in the sample, an r of .632 is needed to justify the assumption that the chances of better than zero correlation are as much as 19 out of 20 (the 5 per cent confidence level). Even with a sample as large as 100 cases, an r of .197 is required to reach the 5 per cent level or one of .256 for the 1 per cent level of confidence according to the t test, as compared to .164 and .233 when the older but erroneous method is used.

A question that often arises in connection with a program of testing has to do with the discriminative ability of a given test. How effectively will it classify individuals with a stated range of talent? If a reading test is needed which will classify fifth-grade children into achievement groups on the basis of their skill in reading, a test that is only capable of making distinctions between the average reading ability of fourth- and eighth-grade children will not serve. Just what the self-correlation or the correlation with some outside criterion should be in order to meet the requirement specified is not an easy question to answer since outside criteria are unlikely to be perfect measures, nor is the test itself likely to be one that measures all aspects of reading ability. Since the standard error of estimating a criterion score on the basis of a single test score is equal to $S.D. \sqrt{1 - r^2}$, where $S.D.$ = the standard deviation of the distribution of test scores and r is the correlation of the test with the criterion, it follows that an r of .866 between test and criterion is needed if the standard error of estimate is not to exceed one half of a standard deviation of the distribution of scores within the range of talent for which the distinctions are to be made, in this case, among children in the fifth school grade.¹³ If the criterion used is reasonably satisfactory, a three-way classification of the subjects as good, average, and poor readers could be made on the basis of such a test with small likelihood of serious error.

Let us now suppose that in the search for a test that will meet this requirement two are found, the first of which has a correlation with the criterion of .92 while that of the second test is but .80. Each of these correlations is based upon a representative sampling of 84 fifth-grade cases. The second test, however, is to be preferred on the basis of its lower cost. What is the likelihood that if calculated for all the fifth-grade children in the city, its true correlation with the criterion would be as high as .866?

Since r is not normally distributed we shall first transmute the two values—the obtained r of .80 and the required r of .866—into their

¹³ It should go without saying that the correlation should be computed on the basis of a representative sampling of children within the range for which the distinctions are required.

equivalents expressed in terms of Fisher's z function. Reference to Lindquist's table shows these figures to be, respectively, 1.098610 and 1.313500. The difference is .214890. The standard error of the z function is $1/\sqrt{84-3} = .1111$. The ratio of the difference to its standard error is 1.934. Reference to the probability integral shows that there are fewer than three chances in one hundred that the second test is truly capable of making the discriminations called for. On the other hand, there are approximately twelve chances in a hundred that the correlation in the universe is as low as .70 while the likelihood that the true correlation may not be above .65 is very nearly as great as that it may be as high as the required level of .866. Because the sampling distribution of r is so markedly skewed at the higher levels, the chances that the r obtained for a second sample will be higher than that obtained for a first sample are by no means equal to the chances that it will be lower than the first, provided always that the first r is not zero. The further the r first obtained departs from zero, the more likely it is that a second sample will show a lower value and the fewer are the chances that its r will be higher than that first obtained.

One might also wish to consider the chances that the other test would *not* meet the requirement laid down for the population as a whole since the correlation of .92 obtained for the first sample may not appear to have a sufficient margin of safety when the findings for the second are taken into account. The equivalent of an r of .92 in terms of the z function is 1.589033. Following the same procedure as before, we find that there is less than one chance in a hundred that this test would not serve the purpose, that is, that the correlation obtained for other samples would fall as low as .866 even though the chances are against their reaching the level of that first obtained.

Like all statistics, the r obtained for a sample affords only a measure of probability as to its value in the universe from which the sample was drawn. Unlike most statistics, however, its sampling distribution is neither normal nor symmetrical but is skewed in the direction of zero. This means that in determining the likelihood of divergences from the obtained value of r in other samples from the same universe, separate calculations must be made as to the probability of increase or decrease since the latter will be greater than the former in all cases where r exceeds zero. Failure to take account of this difference in chance expectation has led to many disappointing results in the field of mental measurement. It is well to remember that when the "goodness" of a test is to be judged on the basis of either its self-correlation or its correlation with other criteria, conservatism rather than undue optimism is the safest attitude to adopt.

Analysis of Variance

BASIC CONCEPTS

The variance of a series of measures is the mean of the squares of the deviations of the individual measures from the mean of the group. In other words, it is the square of the standard deviation.

In previous chapters it was shown (1) that a knowledge of the variation of the measures within a sample makes it possible to predict with a specified degree of assurance, the probable variation of the means of additional samples drawn from the same universe. The standard deviation of the means of this hypothetical series of samples is known as the *standard error of the mean*. It indicates the amount of variation in the means of successive groups that may fairly be attributed to the chances of sampling or, conversely, the probability that the mean of a later sample will diverge from that found for the first by any stated amount if each is truly a random drawing from the same universe. Extension of this principle to a comparison of the findings for two samples leads to (2) the *t* test, which enables us to state the likelihood that the universe from which the second sample was drawn is *not* the same as that from which the first was taken. Not only means but other statistics may be compared in the same way by the use of the appropriate formulas.

The assumption here is that the variance of the means of samples (designated in most texts as the "variance between groups") will not differ from the average variance of the individual cases within the samples (the "variance within groups") to a degree greater than that to be expected by chance if the groups in question are really samples from the same universe. Two estimates of the variance within the universe are thus possible, one of which is based upon the variance of the means of a series of samples, the second upon the mean variance of the cases within the individual samples.

It is, of course, unlikely that the two estimates will show exact agreement with each other, but their divergence may be expected to follow the usual rules of probability. The variance ratio (the *F* test) may accord-

ingly be used to determine the likelihood that the number of times by which the variance of the means of the samples exceeds the mean variance of the individual measures within the samples would occur by chance if all the samples had been drawn from the same universe, or, what amounts to the same thing, from universes that did not differ from each other with respect to the measurement in question.¹ If F proves to be larger (indicating a greater probability of discrepancy) than that which the experimenter has decided upon as the maximum risk he is willing to take, the hypothesis that the universes from which the samples were taken are similar with respect to the measurement in question will be rejected with the appropriate degree of confidence.

PRACTICAL APPLICATIONS

Methods for the analysis of variance were originally devised by R. A. Fisher for use in certain agricultural experiments in which control of conditions could be made much more exact and detailed than is usually possible with human beings. The requirement that the groups used be random samples of the universes from which they were drawn is basic but one that is frequently very difficult to meet because of the complexity of human behavior and the many interrelated factors that

¹ The principle is not difficult to understand. Suppose that we have five universes, each represented by 100 cards numbered to indicate their value. The values of the first set range from 1 to 10; the second has a range from 5 to 15; the third from 10 to 20; the fourth from 15 to 25; and the fifth from 20 to 30. Now if we make a single drawing of 10 cards from each universe and find the mean and the variance of each of these 5 samples separately as well as that for the total when all the samples are combined, two things immediately become apparent. First, the variance of the total is much greater than that of any one of the samples taken by itself or than the mean of all the sample variances. In the second place, the variance of the means of the samples (note that this is not the same thing as the mean variance of the samples) is likewise greater than the variance within the separate samples. This necessarily follows from the wide difference in the range of scores included within the universes from which the samples were drawn. But if the five universes had all been exactly alike, it obviously would make no difference whether all the samples had been drawn from a single one of them (of course returning each sample to the pool and mixing it with the others before drawing the next), or whether each universe had contributed a sample as in the first example given. The estimate of the variance of the total from the variance of the means of the samples drawn from the five groups of which the total is made up would not, in this case, differ significantly from that based upon the mean variance of the five samples taken from a single group.

The effect of combining several groups with unequal means into a single group will thus be an increase in the variance of the total over that of the individual groups. The ratio of the mean variance of the samples (the variance within groups) to that of the means of the samples (the variance between groups) will therefore make it possible to say how much warrant there is for rejecting the hypothesis that the universes from which the samples have been drawn do not differ from each other with respect to the measurement under consideration.

underlie its manifestations. It is true that these are matters which affect the dependability of results obtained by other methods as well as those secured through an analysis of the variance. The errors introduced when the implicit assumptions of the older methods of statistical treatment are not fulfilled are presumably quite as great as those that result when the conditions requisite for the valid use of the variance methods have not been met. The latter, however, are on the whole somewhat more rigid than the former and for this reason the method is used chiefly in experiments formally set up when the effect of some interposed condition is to be tested by a comparison of small and highly selected groups chosen from universes of known composition by the use of random numbers or by some other recognized device for ensuring randomness. In mental testing the method finds its chief applications in the study of the effect upon test scores of such factors as special coaching or other forms of training in which the groups to be compared are drawn at random from the same universe *before* the interposition of the special condition and are retested at the end of the training period. If but a single method is to be used, the t test will provide the most valid answer to the question at issue, but if the effect of several different methods is to be tested simultaneously, with a view to ascertaining the stability of the test results under a variety of external conditions, the variance method is not only more exact but also more efficient since it permits the simultaneous comparison of all the interposed conditions by means of a single operation. However, it must not be supposed that if, when the F test is applied to data of this kind, the variance between groups proves to be significantly greater than the variance within groups, all or any particular one of the methods used has produced a significant change in the test results. If six different conditions have been tried, each with a separate random sample, only one of the six may show results that depart significantly from the generality. Or three may have induced changes while the other three may have had no reliably determined effect. Of the three that proved effective, one may have had a dependably greater influence than either of the others. The last fact, however, is not given directly either by the F test or by the t test, which merely indicate the likelihood that *some* difference exists. It may, however, be determined by a calculation of fiducial limits for which a method will be described in a later section of this chapter.

If it is desired to ascertain which of the various groups differ significantly from each other, a separate comparison becomes necessary for each pair. Since this is the case, one may ask if it would not have been better to apply the t test to the separate pairs in the first place if information about the effect of the individual conditions is desired. Has the

analysis of variance yielded any information that could not have been obtained more conveniently by other methods? Why not, for example, begin by comparing the two methods, which, by surface inspection, appear to be most and least effective and proceed from there to a point at which the differences become so small as to be negligible?

The answer is twofold. First, the procedure last mentioned violates the rules of random sampling since it involves a selection on the basis of facts that could not have been known at the outset of the experiment. Second, it runs contrary to the common-sense principle that if one wishes to make the best possible estimate of an unknown fact, he should make use of all the available data that have a bearing on the matter.

The first point may be illustrated as follows. Suppose we wish to ascertain whether or not the scores on a certain brief test of mechanical aptitude obtained at the time of college entrance vary significantly with the type of secondary school that had been attended. Ten schools are selected for study, including a large city high school, a consolidated rural high school, a small private preparatory school, a public high school in a small town, a large and well-known private preparatory school, and so on. All the entering freshmen from each of these schools are catalogued and a random selection of ten cases from each is made according to one of the methods described in Chapter 8. The score made by each student in the ten samples is shown in Table 10.²

The results of two estimates of the variance of scores of a freshman population made up of an equal number from each of the ten schools according to the data of Table 10 are shown in Table 11. The first estimate is based on the variance between schools; the second on that within schools.

The first column (headed d.f.) shows the number of degrees of

² No attempt will be made here to present more than a brief working outline of a few of the more simple and elementary features of the variance method, since an adequate account of the procedures would require far more space than can be allotted to the topic here. Those who wish to gain command of the method should consult the very complete and detailed accounts given by R. A. Fisher (1936, 1937), Lindquist (1940), Rider (1939), or Snedecor (1946). All that will be attempted in this chapter is to indicate a few of the many advantages of the method for the solution of certain problems and to call attention to some of the misconceptions and limitations in its interpretation that are sometimes overlooked by those whose enthusiasm for this new and undoubtedly very powerful tool has led them to lose sight of its basic assumptions. There can be no doubt that the variance method permits a more effective use of small samples than has hitherto been known since the method makes it possible to check each of a series against all the rest in one simultaneous operation. By thus increasing the number of comparisons within the data, greater precision is gained. But it must not be forgotten that random choice of the subjects comprising the samples is an essential feature of the method, and this is not always easy to achieve when human subjects are used.

TABLE 10
DISTRIBUTION OF SCORES ON A TEST OF MECHANICAL ABILITY
BY STUDENTS FROM DIFFERENT HIGH SCHOOLS

Students	High Schools									
	I	II	III	IV	V	VI	VII	VIII	IX	X
1	4	5	5	2	1	2	4	3	4	3
2	3	1	4	5	3	2	5	3	5	2
3	4	4	4	5	3	3	2	3	4	4
4	2	4	5	5	5	5	5	5	3	5
5	5	3	2	4	3	3	4	2	3	5
6	4	3	1	3	2	1	3	5	2	2
7	3	3	5	3	5	5	2	2	5	3
8	3	4	3	2	1	4	1	2	1	2
9	2	4	3	5	1	1	5	1	2	3
10	4	2	3	5	5	4	5	4	5	3
TOTALS	34	33	35	39	29	30	36	30	34	32
MEAN	3.4	3.3	3.5	3.9	2.9	3.0	3.6	3.0	3.4	3.2
GRAND TOTAL	332									
GENERAL MEAN	3.32									

TABLE 11
ANALYSIS OF VARIANCE FROM DATA OF TABLE 10

Source of Variance	d.f.	Sum of Squares	Variance
Between high schools	9	8.56	0.95
Within high schools	90	163.20	1.81
Totals	99	171.76	1.73

freedom for each estimate. Since there are ten schools, the first estimate, which is based on the means of schools, will have $10 - 1 = 9$ d.f. Since there were ten students chosen from each school, each of these samples would also have 9 d.f. For the total of ten schools, this makes 90 d.f. for the estimate based on the total variance within the separate schools. The total number of degrees of freedom is 1 less than the total number of cases, that is, 99, which, it will be noted, corresponds to the sum of the two components. This is quite as it should be.

The column headed *Sum of squares* is the sum of the squares of the deviations of the measures upon which the different estimates of the population variance are based, taken from the group mean. In practice this is most easily obtained by finding the sum of all the measures, squaring it, and dividing by the number of cases and treating the result

as a "correction term" which is to be subtracted from the sums of the squares of the other terms in order to obtain the divergences. This is equivalent to making all the subtractions at once instead of performing each separately. In the example given, the sum of the 100 measures is 332. The square of 332 is 110,224 which, after dividing by 100, gives 1102.24 as the correction term. The sum of the squares for the variance between means of schools is found by squaring the sums of the measures in each school separately, that is, the sums of the columns, dividing by the number of cases in each school (in this case, 10), and subtracting the correction term. The variance within schools is commonly found by subtracting the variance between schools from the total, but it may also be computed separately to serve as a check on the accuracy of the arithmetic. To find the sum of squares, the individual measures in each school are squared, and from the sum of the squares is subtracted the square of the sum of the measures divided by the number of cases (10). This is done separately for each school. The sum of the ten differences is the sum of squares within schools. The total sum of squares, it will be noted, is equal to the combined sums within and between schools. It can also be computed directly by finding the sum of the squares of all the individual measures in the ten schools and subtracting the correction term.

The third column, headed *Variance*, gives the three estimates of the variance within the universe from which the samples were drawn. It is obtained by dividing the sum of squares for each estimate by the number of degrees of freedom for that estimate. The significant comparison to be made here is that between the variance as estimated from the means of the different schools (the variance between groups) and that estimated from the individual measures within the schools (the variance within groups). If the former is dependably greater than the latter it means that there are differences in the average performance of students coming from different high schools that are too great to be reasonably accounted for by chance differences in the sampling. In the example given, however, the variance between schools is actually *less* than that within the schools, less, that is, than was to be expected on the basis of the individual differences within that part of the college population from which the ten samples were chosen. There is accordingly no reason whatever for rejecting the hypothesis that scores on the test in question are unrelated to the type of school in which preparation for college entrance was made.³ It is accordingly not only unnecessary to make

³ Note that this is *not* equivalent to saying that no such relationship exists for the schools in question, still less that a relationship might not have been found if other schools, not included among the samples, had been considered.

further comparisons between the schools but to do so would, as was indicated in an earlier paragraph of this section, constitute a serious violation of the principles of random sampling, as well as those of common sense.

Suppose, for example, we were to measure the heights of a representative group of one hundred men, all of the same age, and all coming from middle-class white American stock. The distribution of the measures would presumably conform fairly closely to that of the normal curve. Now if we were to *select* the five tallest and the five shortest members of the group and treat these selected cases as if they had been chosen *at random* from two different populations whose ranges of heights were unknown, we would in all probability be led to the conclusion that the groups in question really differed in height, since the most likely assumption that could be made from such data would be that the five specimens chosen in each group came from near the midpoint of their respective populations where the number of cases, and the consequent likelihood that an individual will be chosen, is greatest. Actually, however, these cases are *known* to have been taken from the extremes. To ignore this knowledge would be contrary both to common sense and to statistical theory. It would certainly be as foolish to discard the information provided by the measurements of the other ninety cases as it would be for a doctor to base his diagnosis upon one or two symptoms only, disregarding others of equal importance for the determination of the disease.

Returning now to the question raised earlier: Why not *select* the two schools for which the means of the samples show the greatest difference, viz., Schools IV and V with means of 3.9 and 2.9, respectively, and by applying the *t* test ascertain the likelihood that the students coming from these schools really do differ in their performance on the test in question? In the absence of other information this would be a legitimate procedure that, in this particular instance, would give a value of *t* equal to 2.25 which, for 18 degrees of freedom $(10 - 1) + (10 - 1)$, is significant at slightly better than the 5 per cent level of confidence. Roughly speaking, we may say that there would be about 4 chances in 100 of securing a difference of this magnitude *if the hypothesis of random sampling had been fulfilled*. We should then have been inclined to reject with a moderate degree of assurance the hypothesis that the two schools were samples of the same population. This affords a concrete illustration of the fact that probability is by no means certainty, and that any sample *may* be an extreme case in the population to which it belongs, although the chances are against it if no selection has taken place to bias the results.

One of the great advantages of the variance method lies in the fact that it enables us to judge the significance of the difference between any two measures of a series in terms of all the others. As long as we were able to make comparison between only two groups at a time, our estimates of significance were less precise; by the use of the variance method a considerable gain in exactness of comparison, not only for the series as a whole but for the parts of which it is made up, can be brought about.

When the variance between schools was divided by the variance within schools, what if the ratio obtained had proved to be large enough to meet the requirements likely to be set for rejecting the null hypothesis, say as large as 5.0? We should then be warranted in assuming that differences between schools, too great to be accounted for by the chances we are willing to take, do exist, but this does not mean that each school is different from every other in this respect. A convenient way of classifying the schools with respect to the relative standing of their students on the test in question is to ascertain how great the difference in mean scores must be to meet a specified level of confidence. This makes it possible to ascertain with what level of confidence one may assume that any two schools do not belong to the same universe by direct comparison of their means. For such a comparison, the number of degrees of freedom is the same as that within groups, in this case, 90. Our task is to determine the amount of difference between the means which would warrant rejecting the null hypothesis with any specified degree of confidence. In other words, we wish to ascertain the *fiducial limits* of the difference.

When there is available a series of samples instead of just one, our best estimate of the population variance (assuming that the variance of the several groups differs only within the limits of chance) is the variance within groups, and its square root will accordingly be the best measure of the standard deviation of any one of the groups. The best estimate of the standard error of the mean will be had by dividing the standard deviation by the square root of the number of cases in the group. Following this procedure with the data of Table 10 we have $1.81/\sqrt{10} = .572$ as our best estimate of the standard error of the mean of a single group. Assuming equal variance for the groups, the standard error of the difference between any two groups thus becomes $\sqrt{\sigma_M^2 + \sigma_M^2} = \sqrt{2(.572^2)} = \sqrt{.654368} = .809$. This gives us a basis for ascertaining how great the difference between two groups must be for any given level of confidence. With 90 degrees of freedom, the value of t at the 1 per cent level is 2.635; at the 5 per cent level it is 1.99. Hence the actual differences must be at least $.809(2.65) = 2.132$ to reach the 1 per cent level, or $.809(1.99) = 1.610$ for the 5 per cent level. It will be recalled that when the samples from Schools IV and V were compared on the assumption

of random selection, the null hypothesis was rejected with approximately a 4 per cent level of confidence. But when appropriate allowance is made for the known fact that the choice of schools was not made at random but because they represented the two extremes of a series, the difference of 1.00 in their means is found to fall considerably short, even of the 5 per cent level.⁴ This illustrates how serious an error would be involved by ignoring the evidence provided by the other eight schools in the group, that is, the position occupied by the two schools within the series as a whole. The point has been stressed because it is so frequently ignored in experiments on group differences. There is no doubt whatever but that many of the "significant differences" reported in the literature would lose much of their "significance" if the proper statistical methods had been used in evaluating them.

A second example of the use of the variance method in the testing of a scientific hypothesis is shown in Tables 12 and 13. The data are taken from an unpublished study by the writer. Table 12 lists the scores made on a masculinity-femininity key derived from the differences in the associative responses given by male and female subjects to a list of 238 common words. Four hundred persons of each sex were used in developing and standardizing this key. In the present instance, the hypothesis to be tested is that married, divorced, and single women do not differ in respect to the masculinity or femininity of their responses to this list of words. Samples of ten cases each were drawn at random from larger groups classified on the basis of their marital status at the time of taking the test. The method of scoring is such that positive values indicate a preponderance of responses characteristic of males rather than of females while negative scores mean that the greater number of the responses are similar to those usually given by women. An analysis of the variance was made according to the same method as that used in the example given on pages 274-276. The results are shown in Table 13.

The chief thing to be noted in Table 13 is the large difference between the variance of the group means and that of the individuals making up each of these groups. The F ratio (9886.3 divided by 2642.6) is 3.74. By consulting the table of F values it will be seen that with 2 and 27 degrees of freedom, F should equal 5.49 to reach the 1 per cent level of confidence, but that a value of 3.35 is sufficient to meet the requirements for the 5 per cent level. In the present example the obtained figure (3.74) is greater than this. The hypothesis that the three marital groups do not differ with respect to the masculinity or femininity of their associative responses to this list of words is thus refuted with a moderate degree of confidence. Inspection of the scores indicates that the difference

⁴ Roughly, the difference falls at about the 25 per cent level.

TABLE 12
DISTRIBUTION OF SCORES MADE BY WOMEN OF DIFFERING
MARITAL STATUS ON A TEST OF MENTAL
MASCULINITY-FEMININITY

Married Women	Divorced Women	Single Women
-153	+24	- 75
-108	-33	+ 10
- 44	-11	- 66
- 68	-68	- 79
- 36	+23	- 2
- 22	- 7	- 27
+ 4	+17	- 53
- 24	+34	-185
- 53	+25	+ 38
-117	-50	-115
Totals -621	-46	-554
Means -62.1	-4.6	-55.4

TABLE 13
ANALYSIS OF VARIANCE

	d.f.	Sum of Squares	Variance
Between groups	2	19772.6	9886.3
Within groups	27	71351.7	2642.6
Total	29	91124.3	3141.9

$$F = 9886.3 \div 2642.6 = 3.74$$

is due for the most part to the greater masculinity of the scores made by the divorced women. This finding was confirmed when scores from the entire group of 300 cases were tabulated. Terman and Bottenwieser (1935) obtained similar results from the use of the masculinity-femininity key of the Strong Vocational Interest Blank.

The method described here illustrates only the simplest type of analysis of variance. Extension of the procedure will in many cases indicate the presence of a third component, sometimes called the "remainder" but more frequently known as the "interaction" component. For example, suppose it is desired to compare the effectiveness of several different methods of teaching fifth-grade children to handle fractions. A first experiment in which only a single class is taught by each method indicates fairly marked differences in the mean performance of the different classes at the end of the period of instruction, with the variance ratio exceeding the 1 per cent level of confidence. It is pointed out,

however, that the differences may not indicate superiority of certain methods at all but may only reflect differences in the teaching ability of the various teachers.⁵ Duplicate experiments with several classes assigned to each method instead of just one, as well as a somewhat more elaborate method of analyzing the data, may sometimes produce a relatively straightforward answer to the question, with Method A consistently and reliably superior to Methods B and C, which are about equally effective, while Method D proves definitely poorer than any of the others. Often, however, it will be found that the two components previously considered—the variance between methods and the variance of the children within a single “methods group”—do not completely account for the total variance when all the cases are combined. The remainder, if greater than can reasonably be accounted for on the basis of chance, is most likely to have arisen from an interaction between the kind of method used and the situation in which it is used. From the standpoint of general teaching ability, for example, Teacher A and Teacher B may rank equally high, yet their habitual ways of accomplishing their results may be very different. It may thus come about that Method 1 (generally a good method) may be well suited to the usual practices of Teacher A, who will therefore obtain good results by its use, while Teacher B who, when free to follow her own methods, equals or even surpasses Teacher A in the classroom, finds Method 1 so different from her accustomed procedures that she is very inept in using it. Thus the variance in the total may be due not only to the variance within and between the subgroups of which the total is composed but also to the relation between these components, to the fact that there may be circumstances which cause the members of the various groups to react differentially to the factors on the basis of which the groups were divided.

For an account of the methods used in carrying out more elaborate investigations such as that just mentioned the reader should consult the references cited on page 274. Before passing on to the next topic, however, a few words of caution should be added. In the first place, the undoubtedly greater precision of these methods when properly used cannot make up for failure to maintain the controls necessitated by the basic assumptions from which the statistical formulas have been developed. The methods assume that the subjects within each sample group to be compared have been chosen *at random* from the groups in question. If the problem consists in testing the effect of some condition or conditions

⁵ In all experiments of this kind it is assumed that the assignment of pupils to the different classes has been randomized in some way. Otherwise, any difference found might be due to differences in the intellectual level of the children rather than to differences in the manner in which they are taught.

experimentally introduced, then either homogeneity of variance must be assured by placing an equal number of subjects of each initial level of competence within the several groups *before* the experiment is started, or by a carefully devised plan for securing a random assignment of subjects to the different groups. The first plan involves correlation between the groups and this must be allowed for in determining probability levels. The second involves access to a sufficiently large universe so that the withdrawing of one sample will not affect the composition of those chosen later to too serious an extent. In the agricultural experiments for which these methods were originally devised, the requirements of randomness and homogeneity of variance from group to group were not difficult to fulfill and, while recognized, were not greatly stressed. If, for example, an experiment on the effect of different fertilizers upon a certain strain of wheat is to be tested, it is by no means difficult to secure a series of samples of wheat on which the different fertilizers are to be used which will conform very closely to the required conditions of random selection and approximately equal variance. But when human beings are substituted for grains of wheat, one cannot depend upon the relatively simple methods found serviceable in agricultural research. The source of supply from which the samples are drawn is likely to be much more limited, its parts may be unequally accessible, the variance of the universe may therefore be poorly represented by the variance of the combined samples which is taken as an estimate of it, and the means of the samples may differ from each other for causes very different from those which presumably formed the basis for the separation.

For these and other reasons, the neatly designed experiments with a limited number of cases which have proved so effective in areas where the selection and organization of the research material is almost if not wholly under the control of the experimenter are not easy to duplicate with human subjects. The agricultural research worker also has the very great advantage of knowing the history and characteristics of the subjects with which he works in a way that is rarely if ever possible when human beings are concerned.

All this makes it possible to set up investigations on plants or animals which provide models of technical precision that the psychologist may emulate but will rarely parallel. He may copy the form, and if he takes all possible care to meet the required assumptions as far as this is feasible, he may achieve a distinctly worth-while gain in efficiency of procedure and in the confidence warranted by the results. It is well to remember, however, that the human being is a more complex organism than the soya bean or the dairy cow, that his habits and ways of living are self-determined, while those of the domesticated plant or animal are

imposed from without.⁶ This always introduces the possibility that factors other than those in which the experimenter was interested may have been the chief agents in bringing about the results obtained. One may separate the total variance in a given population into its component parts with a high degree of assurance and yet be woefully in error when attempting to ascribe either the variance between groups or that within groups to some particular condition or circumstance. The fact that causation cannot be inferred from correlation is now pretty generally understood, but the terminology employed in the analysis of variance has led many to overlook the fact that the same limitation holds for the new as for the older methods.

Agricultural workers very commonly regard the 5 per cent level of confidence as satisfactory for their purposes. Psychologists may find it wise to be somewhat more conservative than this because of the greater number of factors likely to affect the results of their experiments that are beyond their power to control. It should be remembered, however, that conservatism does not always lie along the road of failing to reject the null hypothesis. There are cases in which failure to reject the hypothesis when such rejection is truly called for may lead to more serious error than would result from rejecting it when later developments show that the obtained differences were in all probability due only to chance. Questions such as these depend upon the particular problem and must be answered by the individual investigator. No general rule is possible.

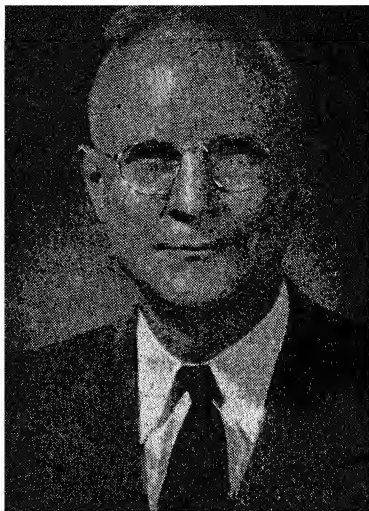
THE DESIGN OF EXPERIMENTS

One of the most brilliant contributions of R. A. Fisher and his colleagues is their demonstration of the tremendous gain in efficiency that can be brought about by designing an experiment in such a way that every part of the data may yield the maximum amount of information possible and all unnecessary duplication may be avoided. For example, it might be desired to test the effect of six different fertilizers on as many different varieties of potato. In a plot of uniform soil and exposure to sunlight would be planted one row of each of the six kinds of potato, each row containing six hills. In the rows running at right angles to these the six kinds of fertilizer would be applied, one kind in

⁶ This statement, of course, must not be taken too literally. Neither man nor animal lives free of external constraint, nor can either develop except in accordance with the laws of his own nature. But it is unquestionably true that the behavior of human beings varies according to principles that are less completely understood than those governing the behavior of organisms below the human level. Consequently, human behavior is harder to predict and control.



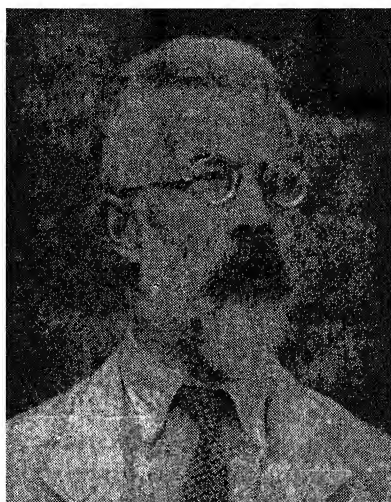
TRUMAN LEE KELLEY



J. P. GUILFORD



L. L. THURSTONE



R. A. FISHER

FIG. 18. SOME OF THE STATISTICIANS WHO HAVE MADE NOTABLE CONTRIBUTIONS TO THE FIELD OF TESTS AND MEASUREMENTS.

each row. In this way, every variety of potato would have been subjected to each kind of fertilizer with other conditions kept the same for all. As many duplicate experiments could be made as was deemed necessary, either under the same conditions or in plots where soil or other factors were intentionally made different from that of the first experiment.

Many modifications of this relatively simple design are possible. An important application of the principle to the field of mental testing is seen when it is desired to study *simultaneously* the effect upon test scores of a fairly large number of different factors such as sex, socioeconomic background, rural or urban residence, amount of schooling, and so on. The method used is that of the so-called *Latin Square*. The experimental setup differs from that of the potato-fertilizer experiment described above only to the extent that instead of having each row and each column devoted to one and only one kind of potato or fertilizer, the various factors to be considered are randomized in such a manner that each enters into the total an equal number of times in a relationship so planned as to give each factor equal weight in the total. An example of a five-variable design is given below:

A	B	C	D	E
B	A	E	C	D
C	D	A	E	B
D	E	B	A	C
E	C	D	B	A

Each letter represents a single factor; it will be noted that each is found once and only once in each row and in each column. The procedure for setting up such an experiment and for analyzing the results is fully described in Fisher's *The design of experiments* and will not be repeated here. Additional notes on the method as well as very complete series of Latin-Square designs will be found in the *Statistical tables* by Fisher and Yates. Lindquist (1940) suggests a number of applications of the method to the field of educational research, especially to those questions that have to do with the comparison of different methods of instruction.

Some Questions of Mental Organization

EXPLICIT AND IMPLICIT HYPOTHESES

Everyone, whether he is aware of it or not, has formed certain concepts with respect to the way the mind works. A few have attempted to formulate their ideas in terms of explicit theories regarding the structure of mind and have developed a number of ingenious ways for putting their ideas to objective test. Preceding chapters have mentioned some of these theories and the procedures used for verifying them. For the most part, however, these studies have dealt only with the "whats" of behavior with but scant attention to the "hows" or the "whys." This was perhaps inevitable in a field where so little was known, for assuredly not much is to be gained by attempting to explain an event until enough is known about its main characteristics to provide a reasonably clear idea of what is to be explained.

Until the beginning of the present century, attempts at developing a consistent and coherent theory of mental organization can best be described as logical rather than psychological. There was the point of view that mental acts are of three kinds—cognition, conation, and feeling—for which James Ward was primarily responsible, but which was later taken up and modified to some extent by George F. Stout. Neither Ward nor Stout was an experimentalist; their interests lay in drawing up a closely organized system into which common observations could be fitted without undue violence. There was the "faculty psychology" which we have mentioned before. There were many other attempts to organize the observed and recorded mental activities of man into a unified structure in which all the parts would conform to the basic design. But it was not until 1904, when Spearman published his first exposition of the two-factor theory, that a serious attempt was made to apply mathematical procedures to mental measurements for the express purpose of developing and testing a complete theory of mental organization. We have shown in a previous chapter how Spearman's theory grew out of his studies of tetrad differences. The uniformity of the cross

products in the correlation matrixes he drew up, which, it will be recalled, indicates that the correlations in question can be accounted for by a single common factor, led him to the belief that this factor is the same for all mental activities. For if a single factor can account for the interrelations of variables *a*, *b*, *c*, and *d* and also for those of *c*, *d*, *e*, and *f*, and so on, as new factors are added indefinitely, the identity of the underlying factor from series to series seems adequately proven. The fact that the correlations always fall considerably short of 1.00, however, indicates that other factors not common to all the variables enter the picture. The single common factor was designated by Spearman as *g* (general factor), while the others were known as specific factors (*s* factors). Spearman held that *g* plays a part in every mental act, although some acts are more dependent upon it than others. He also held that individuals differ in respect to the amount or strength of the *g* which they are able to bring to bear upon the tasks they attempt to perform. The difference between the genius and the idiot can be largely described in terms of their relative endowment with *g*.

The *s* factors are many and various. Some enter into the performance of a large variety of tasks of a given type; they are sometimes called group factors. The scope of the group factors may be fairly broad but they always lack the complete generality of *g*. Others are highly specific, playing a part in the performance of only a small number of mental acts. As is true with *g*, individuals differ with respect to the kind and the strength of their *s* factors, and this accounts for the wide differences in the patterns of mental ability displayed by different people, for the special talents and deficiencies that may and do appear at practically every level of general intelligence. In *The abilities of man* (1927) Spearman also suggested that there may be other general factors which, like *g*, enter into all mental acts but which differ from *g* in that they have to do with its mode of operation, rather than with its amount or strength. To the two for which Spearman was able to find the most convincing evidence he gave the names of *c* and *w*. In popular language, *c* has to do with freedom from inertia, quickness of thought processes. *W* signifies will power, self-control, the ability to persist in the face of difficulties without oscillation. Spearman believed that an adequate measurement of these three general factors, together with a sufficient number of the most important *s* factors, would provide the basis for a truly scientific description of human abilities which would serve most practical purposes as well.

One of the most active opponents of Spearman's theories was E. L. Thorndike. Not only did Thorndike criticize Spearman's theories on statistical grounds, pointing out that in many instances the self-corre-

lations as well as the intercorrelations of the measures used were too low to provide the kind of evidence needed, but he also questioned the very existence of such a universal trait as *g*. According to Thorndike, we have no "intelligence" but only a very large number of "intelligences" made up of elements that overlap to varying degrees. Generality, insofar as it exists, is resident in the nature of the acts performed and only secondarily in the person who performs them. Thus, certain acts call for much the same kind of skill, and within each class of skill are differing levels of difficulty. Correlation between tests or ratings on different traits of the same persons are primarily the result of the fact that, although called by different names, each of the characteristics in question has some features in common with one or more of the others.

In *The measurement of intelligence* Thorndike pointed out that individuals differ in their ability to perform any specified kind of task along at least two axes. They differ in respect to the *number of items* and also in respect to the *difficulty of the items* with which they are able to deal successfully. Thus we find persons who know a great many elementary facts about a wide range of different topics but not very much about any one of them, while others, whose information along some one line extends far beyond that of most other people, may be amazingly ignorant of things which are commonplace for the majority. Inasmuch as Thorndike, in effect, identified intelligence with skill or knowledge,¹ it is easy to see how cases such as these would militate against his acceptance of the idea of a general factor which underlies all types of performance.

Figure 19 is a schematic representation of the contrasted views of Spearman and Thorndike on intelligence as manifested in behavior. In this diagram each circle stands for the performance of an act or task. It will be noted that a varying portion of each circle is common to all the other circles. According to Spearman's view, only this central portion (marked *g*) is to be regarded as the intellectual component of a given act. Other parts are common to some but not to all of the circles; these are the group factors. The parts marked *s* are specific factors, each of which is peculiar to the particular act or task to which it pertains.

Thorndike's interpretation is different. According to his theory, the interlaced circles represent classes of acts or tasks, each of which is looked upon as a basic and fundamental aspect of mental structure. When used as measures of individual aptitude or skill, scores on the different kinds

¹ As given in the Symposium (1921), Thorndike's definition of intelligence was "the power of good responses from the point of view of truth or fact."

of tasks usually correlate positively with each other because some of their elements are the same. If most of the elements overlap, the correlation will be high; if the overlapping includes only a small proportion of the elements involved, the correlation will be low. If the overlapped portions make up an approximately equal portion of each of two tasks, the correlation between them will show rectilinear regression, but if the

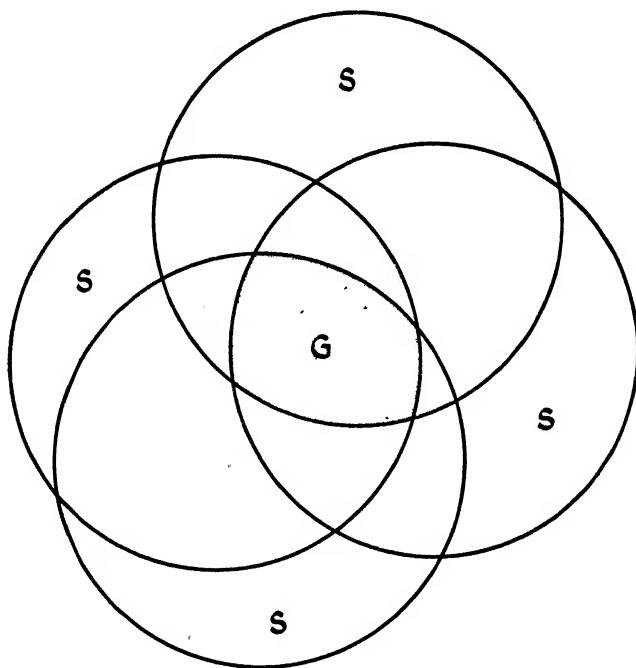


FIG. 19. SCHEMATIC REPRESENTATION OF THE VIEWS OF SPEARMAN AND THORNDIKE ON THE ORGANIZATION OF INTELLIGENCE.

overlapped portion accounts for a larger part of the total area of one task than it does of the other, the regression will be curvilinear and the two correlation ratios will consequently be unequal.

The disagreement in the views of these two eminent men aroused much interest both in the United States and in England and led to many attempts to resolve the controversy through the mathematical treatment of data which seemed better adapted to the purpose than those which Spearman had used. That some modification of Spearman's claims was necessary soon became apparent, since the tetrad difference criterion was so frequently not satisfied. Nevertheless, the concept was a tantalizing one. The possibility of reducing the countless permutations and

combinations of human ability to a relatively simple and clean-cut system of order which could be verified by mathematical treatment fired the imagination and appealed to the scientific spirit of the age. Various methods were proposed, but the keynote of all was the same. All were based on the mathematical treatment of correlation matrixes in the effort to account for the interrelationships of the various measurements in terms of the smallest possible number of factors. Kelley, in his *Crossroads in the mind of man*, held that mental abilities can be reduced to a relatively small number of orthogonal (that is, independent) traits. This is somewhat like Thorndike's view of intelligences rather than intelligence, but Kelley's method of proof was mathematical and, like Spearman's, was based upon an analysis of tables of correlations. Other psychologists and statisticians proposed different theories and new procedures for verifying them, but the general method of factor analysis, as these techniques soon came to be called, became firmly rooted during the decade of the 1920's, and with the passage of time the procedures worked out by L. L. Thurstone have taken precedence over all others. (See Chapter 15.)

Thurstone's view of mental organization comes nearer to that of Spearman than do those of any of the other persons mentioned. His statistical procedures are different from those of his illustrious predecessor but, although he does not find any one factor that can be regarded as completely general, his method does result in an arrangement of the factors which he isolates in order of their generality. Thurstone's first factor may therefore be regarded as somewhat analogous to Spearman's *g* but with this very important difference: Thurstone does not claim that it will remain the same if some of the elements in the correlation matrix are changed or if a different population is used for deriving the correlations. Spearman looked upon *g* as a fixed characteristic of mind, the nature of which is the same for all. Thurstone specifically denies any claim that his interpretation or description of the factors he isolates is the only possible one. In *Multiple factor analysis* (page 332) he states:

To hunt for a unique solution in the comprehension of a set of related phenomena is an illusory hunt for absolutes. It is probably safe to say that an apparently unique set of concepts in any domain is merely the symbol of our immaturity in the exploration of that domain. The recognition of the lack of uniqueness in scientific concepts does not imply that some sets are not more useful and fruitful than others. Those parameters are preferred which reveal the phenomena as of a simple underlying order.

Thus, even the relatively objective views of mental organization that

have been reached through the mathematical treatment of test results are by no means independent of the initial assumptions with which these inquiries were started. These assumptions are in part explicit; their proponents have been aware of them and as a rule have specified them at the outset of their reports. But there are other assumptions, not stated in the reports but too often taken for granted both by the investigator and by his audience. Among these are such questions as the uniformity of meaning of the score variance of different subjects on the same test, whether ability can ever be effectively separated from motivation, and the effects of such factors as Spearman's c and w upon the results of tests and the mathematical analysis of these results. These are matters to which far too little attention has been given; it is to be hoped that in the future their role in mental organization will be overtly recognized.

THE MATHEMATICAL ANALYSIS OF NONINTELLECTUAL TRAITS

That such factors as self-control, level of aspiration, interest and zest in achievement, and a host of other matters by which potential abilities are either energized or constricted in their manifestations play an important part both in performance on mental tests and in the larger problems of real life is generally admitted. In Part III we shall review some of the methods by which various people have sought to reduce the emotional and conative aspects of mental life to objective measurements, and to ascertain some of the external factors which occasion their manifestations. As yet, however, work in this area has not advanced far beyond the level of single measurements; little has been done to show the organization of these nonintellectual traits either with respect to each other or—what may perhaps be more important—with respect to their integration with the abilities and achievements of the individual. Common observation indicates that a basic problem in the field of human behavior is involved in such relationships. How often do we hear such pronouncements as these: "He could if he would, but he won't make the effort." "He's not very clever, to be sure, but he never gives up till he gets there." "He is a good workman but he can't hold a job because of his bad temper." Regardless of the accuracy of the particular statements, the general principle is beyond question.

A number of attempts have been made to apply the methods of factor analysis to a series of intercorrelations of personality measurements, but in general these have met with but mediocre success. One of the difficulties, as Thurstone (1947) has pointed out, is to be found in the fact that in so many cases the social acceptability of a trait does not

vary directly with its amount or intensity. Often maximum acceptability lies at or near the midpoint of the quantitative score, rather than at either extreme. Moreover, such traits are often not clearly defined, and the same trait name may convey very different meanings to different people. All this makes for considerable difficulty in the application of factorial methods to nonintellectual traits. Perhaps the time for such procedures is not yet. It may be that to attempt such analyses with our admittedly very imperfect instruments would be unprofitable, that the task which lies directly ahead is to clarify our definitions and improve our measuring instruments. It may be, however, that if such an analysis were to accomplish nothing more than a first approximation to the task of bringing some degree of order into the 17,953 trait names listed by Allport and Odbert (1936), it would still be well worth while. Cattell (1946) has made an important beginning in this direction in a study which will be discussed in more detail in Chapter 26.

The methods employed in an analysis of the variance have been used but little in studies of mental organization. This might well prove a valuable type of approach, particularly if experiments were set up in such a way as to study the interaction components. Common observation leads to the belief that such factors as motivation, for example, are not fixed characteristics, but vary both with the individual and with the circumstances under which he works. The same conditions will not affect all persons in the same manner or to the same degree. The same individual will react differently when the circumstances are changed. A few well-designed experiments planned with special reference to a study of the interaction between the character of the subjects and the conditions of experiment might well throw new light on the nature of mental organization and on the rules by which human behavior is governed.

No chapter on this subject would be complete without some reference to the topological studies of the brilliant Kurt Lewin. However, Lewin's theories and methods are so radically different from those with which we have been familiar in the past that any attempt to condense them into a few brief paragraphs is likely to be misleading. In general, however, it may be said that Lewin regarded the human organism as an indivisible unity whose behavior conforms to certain fixed rules to which there are no real exceptions. Apparent discrepancies are due to the fact that our understanding of the rules or our acquaintance with the facts of behavior and circumstance is incomplete. Lewin conceived of mental organization in terms of what he designated as "tension systems" or psychical needs by reason of which certain external objects or events acquire an attractive or repulsive character to which Lewin,

borrowing a term from chemistry, gave the name of *valence*. If the valence is negative, the individual tries to escape from the situation; if it is positive, he tries to attain the goal in question or to maintain it after it has been gained. The existence of barriers that inhibit his freedom of movement or the character of obstacles and their position with reference to himself and the goal object are of the utmost importance in determining the behavior. The different tension systems, moreover, are not unrelated to each other but have varying substitute values from which the relative strength of the boundaries between them may be judged. A given psychical need may sometimes be satisfied by the gratification of another, but substitutions of this kind will not always work.

Although Lewin was by no means uninterested in questions of individual differences, he held that these differences merely represent particular instances that can be subsumed under general rules. The character of each can be predicted with assurance if the rule to which it conforms is known. Lewin therefore believed that more is to be learned by setting up controlled experiments aimed at discovering rules than by piling up statistics concerning individual cases without clearly defined aim.

A topological approach² such as that favored by Lewin necessitates many departures from conventional ways of thinking and current ideas of measurement. We no longer ask: How many? or How much? Instead, we must learn to think in terms of movement within psychological space and of the attractions and repulsions, obstacles and barriers by which the character and direction of that movement is determined. The question is not: How many? but In what manner?

In respect to the point last mentioned we may note at least a surface resemblance to the later work of Alfred Binet. It will be recalled that as his work proceeded, Binet inclined more and more strongly to the discovery and description of qualitative differences in the responses of subjects of differing levels of mental maturity when the same problem was set them. A number of these were incorporated into his series of mental tests.³ Since Binet's time, test makers have been less observant of the nonquantitative aspects of test performance. Perhaps Lewin's work may help to awaken a new interest in this area.

²Topology may be defined as the science which deals with nonquantitative relationships within space: with fields of force, directional tendencies, obstacles and barriers, and so on.

³For example, in the tests involving the description of pictures and the definition of common words.

PART III

Tests and Scales

The Conduct of an Examination, with Particular Reference to the Testing of Young or Difficult Children

GENERAL NOTES

No test can be regarded as valid unless it becomes a cooperative enterprise in which the role of the examiner is that of presenting to the subject a standard series of tasks in a manner that will arouse his interest and challenge him to put forth his best efforts. This is a truism frequently stated but often disregarded. Merely going through the form of a test counts for little if the subject's cooperation is poor or if the examiner is careless in his procedures or inaccurate in his scoring. The old saying that "a chain is no stronger than its weakest link" is nowhere more applicable than in mental testing.

Securing the cooperation of school children or of normal adults who present themselves voluntarily for testing, or who readily assent to a request that they serve as subjects in an investigation that involves the use of tests, is unlikely to present serious difficulty to the experienced examiner. The examination of children of preschool age is a different matter. The child of two to six has little awareness of or interest in his own performance as such. His concepts of social behavior are primitive; he is still an individualist who follows his own impulses. He can be allured but not coerced. His likes and dislikes are expressed in prompt and vigorous action; tears, smiles, and temper tantrums all lie very near the surface.

For these reasons the testing of young children necessitates a degree of social perceptiveness that many examiners lack in spite of its importance for clinical work in general. Older children or adults are likely to be relatively patient with the blundering of a poorly trained examiner who fumbles about for misplaced materials, who reads ques-

tions out of a book in a monotonous voice without looking at the subject, and who fails alike to prepare him for what is to come or to acknowledge his responses by so much as a nod or a smile.

The little child will not tolerate such treatment. He demands attention, not fleetingly or occasionally but all the time. If he does not get it his protests are of a kind that cannot be ignored. He snatches material, screams, or runs from the room. He is ready with an unqualified "No," or "I won't" when the task presented does not strike his fancy, or an "I don't like you," for the examiner who makes a similar impression. On the other hand, the happy absorption of a three-year-old who feels at ease with the test and with the examiner, his gleeful anticipation of each new task, and his obvious delight in the examiner's praise is a joy to witness.

It is doubtful whether any other kind of clinical experience is of equal value for training in the interpersonal aspects of testing as is an internship in the mental examination of young children. The situations encountered there do not differ essentially from those that occur in the examination of older persons, but the responses of the small child serve to bring the examiner's blunders out into the open, while the socially acquired inhibitions of the school child or the adult permit them to be overlooked. The following suggestions for the conduct of an examination, although they were prepared as a part of the *Manual of instructions for the Minnesota Preschool Scales*, have wider reference than might at first seem. With slight modification they apply to older children quite as much as to those of preschool age, and particularly so when the subject is nervous or fearful—as is many a child brought to a clinic for examination, or if he is in the angry, rebellious, or suspicious mood frequently encountered among delinquents or children who are school problems. Even in the individual testing of adults, the general principles will be found to hold good.

THE CONDUCT OF AN INDIVIDUAL EXAMINATION OF A YOUNG CHILD¹

Make the testing room attractive to the child. A suitable testing room is well-lighted, adequately ventilated, and not so large as to encourage the child to run about. It contains only the necessary furnishings, but these are attractively arranged. The table and chair to be used by the child are the right height for him and painted a pleasing color. Chairs of several sizes may be kept in an adjoining room and changed

¹ This section is reprinted by permission of the publishers from *Minnesota Preschool Scales: Manual of instructions* by Florence L. Goodenough, Katharine M. Maurer, and M. J. Van Wagenen, Minneapolis, Minn.: Educational Test Bureau, 1940.

when necessary to fit the child. A few pictures which are interesting to children are hung low enough on the walls to permit the child to examine them closely. Furnishings suggestive of a doctor's office are to be avoided, since for many children such surroundings have unpleasant associations.

Make the arrangement of testing materials systematic and habitual. The particular arrangement of testing materials is an individual matter with each examiner, but time is saved and testing proceeds most smoothly if a constant system is adhered to. A plan which has proved satisfactory to many examiners of young children is that of having the child sit at a small table at the left of the examiner.² The examiner sits at a higher table with the materials in a box at her right or at a single pedestaled desk with the materials in the drawers at her right. This arrangement has the advantage of putting the examiner between the testing materials and the child, and of placing the testing materials out of reach and out of sight of the child when not in use. There are other advantages of this side-to-side arrangement of tester and child. Young restless children need help in keeping their chairs adjusted and, occasionally, a restraining hand when their fidgeting interferes with the progress of the tests or when they reach for material before the directions have been given. Shy children find it embarrassing to be compelled to meet the examiner's eyes every time they raise their own, as they must if a face-to-face arrangement is used. Very young children are reassured by having the examiner sit close to them. In order to avoid distractions from the succeeding tests, testing materials, toys, and other accessories are removed from the child's table as soon as they have served their purpose.

Keep testing materials, toys, and other necessary equipment always at hand but out of sight. Paper handkerchiefs, paper, pencils, and toys are standard equipment. The toys must be small and selected to stimulate quiet play at the table rather than to encourage the child to get out of his chair. They must not duplicate any of the testing materials. Children are fascinated by tiny articles such as small cups, saucers, and table utensils, wee dolls, and diminutive animals, houses, and trees. Picture books covering a range of little children's interests are also desirable. The toys and books, and the conversation which naturally arises from their inspection, are a great aid in establishing rapport before the tests are presented. A toy or a book may be handed to the child while the tester records an unusually long response during the examination, or when an interruption seems advisable to forestall boredom or fatigue. Paper or a notebook is kept at hand to jot down notes on the child's

² In the case of a left-handed examiner, this position should be reversed.

behavior. These notes are a valuable aid in judging the validity of the examination. The pencils used by the child are to be not more than five inches in length so that they can be easily manipulated by small fingers. Before the child arrives, the examiner should check over all testing materials and accessories to see that everything is in place, and that there is an ample supply of paper, sharpened pencils, and forms for tracing.

Do not urge the child to respond before he is ready. Little children are not used to meeting new people in strange places and must be given time to adjust themselves before they are required to take an active part in a new situation. Nothing is more unfortunate than centering attention on the child immediately upon his arrival.

The inexperienced examiner may make a bad start with the child which is overcome only with great difficulty, if at all, by allowing the mother to insist that the child shake hands with the examiner and say, "Good morning." If the child is reluctant, or refuses to do so, the mother may resort to coaxing, threats, and reproof. Needless to say this does not bring about a friendly attitude toward the examiner. Matters are made worse if the examiner steps in at this point and thrusts toys toward the child who, by this time, may be so bewildered and emotionally disturbed that a storm results. This situation can be avoided by observing the simple, but important, rule: *Let the child make the advances when he is ready.*

The experienced examiner addresses her first remarks to the mother, taking care to distract the mother's attention from the child. After saying casually, "Good morning, Tommy," in a tone which does not require a reply from the child, she immediately turns again to the mother with conversation which has no relation to the child and which definitely excludes him. When the child has had an opportunity to examine his surroundings and appears to be at ease, the examiner calls his attention to the toys, which have been placed on a table a short distance from the mother, and suggests that he may like to look at them and play with them at the table. After a little more conversation with the mother, she is sent from the room and the examiner plays with the child and the toys, encouraging him to respond to her conversation, but not insisting that he do so. When the child appears to be responding freely, the examiner may safely begin the formal tests.

Make sure that the child is physically comfortable before beginning the examination. The examiner suggests to the mother that she attend to the child's toilet needs before the test is begun. All the child's outer wraps are removed, and the temperature of the room regulated with the child's indoor clothing in mind. The examiner has, of course, removed

her wraps, including her hat, before the child arrives. If the child sneezes or shows evidence of having a cold, he is encouraged to use the paper handkerchiefs. The child's chair must be arranged so that the light does not shine in his eyes.

Follow instructions exactly. Slight changes in the wording of questions may have an unsuspected effect upon the child's response. Begin testing at the point specified in the Manual of Instructions. Do not assume success or failure on any required item without trial, nor change the wording of questions that the child does not seem to comprehend.

The limits of testing given in the instructions should be rigidly adhered to. The order of giving the tests is changed only when there is a very definite reason for so doing. A notation stating the change that has been made and the reason for it should be written on the test blank. Since there is undoubtedly some effect of transfer from one kind of test to another, the established order of giving the tests should be maintained so that the effect may be a constant one for all.

Adjust the speed of administering the tests to the personality of the child. Shy and timid children require a slower tempo than active and aggressive children. The examiner, however, must be constantly on the alert for signs of confusion or discouragement which may result from too rapid presentation of material. On the other hand, too slow a procedure may cause boredom or inattentiveness. The aim should be to keep the child an interested and active participant at all times without making him feel hurried.

Keep the voice low in pitch. The examiner should speak distinctly but avoid forced articulation. Children resent being talked down to, whether by the use of babyish expressions or by intonation. The aim should be to use a natural tone, taking care to speak clearly so that directions are understood without strain on the child's part. With excitable or distractible children, an intentional drop in pitch is particularly effective in holding the attention. A pleasing voice and an enthusiastic manner will do a great deal toward maintaining interest.

Prepare the child for each kind of test. Tests quite unrelated in content may be woven into a natural sequence by introducing each new test in an appropriate manner. A comment on the child's last performance such as, "That was fun, wasn't it?" or a hearty, "That was splendid," or (in case the child has not done well and knows it) "That was a hard one, wasn't it?" may be followed by, "Now I have some pictures to show you," or "Here's one you are going to like even better," or "This is an easy one. You will like this one, I know." After a mental set has been established for a certain test, the examiner avoids comments or other unnecessary distractions until the child has completed that series of

items. The skillful examiner does not allow the child to dwell upon past failures. Such failures as cannot be concealed from him are made to seem unimportant. As soon as a test is finished, the child's attention must be directed immediately toward the one that is to follow.

Never ignore a child's remarks. Every remark made by the child should be acknowledged in one way or another, no matter whether this remark is a reply to one of the formal questions included in the test itself or is a spontaneous comment or question unrelated to the test. A nod or a smile will often serve the purpose as well as a verbal response. This does not mean that children should be allowed to chatter indefinitely. The child's attention may be tactfully recalled to the task at hand without ignoring what he says. Children are as sensitive to social situations as adults, and if a child feels that his responses are not of sufficient interest to the examiner to merit a reply, he is likely to meet the situation by refusing to respond to questions asked by the examiner.

Praise adequately. Stereotyped phrases are to be avoided. Such expressions as "All right" given in monotonous succession may irritate the child enough to have an unfortunate effect upon the test score. Facial expression, manner, and tone of voice are just as effective as well-chosen words of approbation. Shy, timid, or retarded children require a greater amount of praise than more confident children. A cock-sure attitude on the part of an overconfident child may be checked by an occasional warning such as, "Be careful now. Be sure to get it just right." Except in the case of an unusually timid child, or a very young one, praise is reserved for the end of a series of tests, in order to avoid giving the child cues as to his performance.

Watch for early signs of boredom, fatigue, physical discomfort, or emotional distress, and do something about them before such conditions become acute. Tests of slight intrinsic interest to children may be made interesting to them by the enthusiasm of the examiner. If the testing period is unusually long, it is sometimes best to interrupt it for a few minutes by taking the child to the toilet if necessary, getting him a drink of water, or at least allowing him to stand up, stretch, and look at the pictures on the walls. If this is done when he shows the first signs of fatigue, he will return refreshed and with renewed interest in the test. When a child shows fatigue near the end of the testing period, he may be encouraged to continue by being told that he is nearly finished. Very small children may need to be taken out to see their mothers once or twice during the testing period, to reassure them. If this procedure is suggested by the examiner before the child has become uneasy, he is quite willing to return.

The test procedure must be so thoroughly learned in advance that the examiner is able to devote herself to the skillful handling of the child. No normal, healthy three-year-old can be expected to remain passively waiting while the examiner looks at a book to find out what she ought to say next. No more than an occasional glance at the instructions should be necessary to refresh the examiner's memory. Inaccurate knowledge of procedure results in departures from the directions or inattention to the child, either of which may invalidate the results.

Be playful and friendly but always maintain control of the situation. The examiner must take care that she does not lose control of the situation in her efforts to see that the child enjoys the tests. There should never be any doubt in the child's mind as to who is the leader in the games, who chooses them, and who decides upon the rules.

Make sure that the child cooperates actively at all times. Merely going through the formal procedure of administration does not constitute a valid test. Even under the best conditions and with the most skillful examiner, an occasional child will be so thoroughly conditioned against new persons and places that it will be impossible to secure his cooperation on his first visit to the laboratory. Because of the unwarranted confidence which is sometimes placed in numerical results, the safest procedure to follow is not to give any test at all, rather than to give a poor test. If such a child is not forced to take the test on his first visit but is allowed to play with the toys at a short distance from his mother, and if precautions are taken to see that he has a good time, he will usually return on a second occasion ready to cooperate actively in all that is required of him. In the case of an unusually shy child, two or three such preliminary visits may be advisable. Pressure of time should never be an excuse for accepting a child's half-hearted compliance with instructions. *It is necessary to keep in mind that it is the child's performance which constitutes the test. The particular combinations of tasks set before him are of little consequence unless the child himself actively applies himself to their solution.*

Summary: Characteristics of a good examiner. The good examiner is alive to the importance of small matters. She is careful of her appearance, avoiding extremes of dress on the one hand and the official severity of a starched white uniform on the other. She maintains an attitude of lively interest in all that the children do and say; is enthusiastic but not "gushy," sympathetic but not sentimental. Her facial expression is mobile; she smiles frequently but does not keep her lips stretched in a meaningless grin. She is socially perceptive, noting and responding to small indications of flagging interest or of physical or emotional distress before they become actively disturbing. She is meticulous about all matters of

test administration and test scoring, and yet is able to weave the separate items into a connected series of tasks that progresses so smoothly and informally that to an outsider the entire procedure seems as spontaneous and unstudied as does the performance of a talented actor when speaking the lines he has so painstakingly memorized previous to his appearance before the footlights.

In dealing with parents and teachers, the examiner should be friendly, courteous, and sympathetic, always ready to treat their opinions with respect, even though they may be greatly at variance with her own. She should be wary of expressing her conclusions in too dogmatic a manner, and should never lose sight of the fact that the results of even the best test are sometimes misleading, particularly in the case of small children, and that it is often better to defer a diagnosis until the earlier findings have been confirmed by tests given at a later stage of development. Every examiner should have sufficient faith in her own competence to be able to say "I don't know."

INDIVIDUAL EXAMINATIONS OF OLDER CHILDREN AND ADULTS

As was noted in an earlier paragraph, although not all the details of the suggestions made in the preceding section are applicable when older subjects are to be tested, the principles involved are the same for all. No matter what the age of the subject, the competent examiner sees to it that he is physically comfortable, emotionally at ease, interested in the test, and eager to do his best with it before the examination is begun. As it progresses, the examiner is quick to note any indications of embarrassment or nervous fatigue on the part of the subject and to reassure him if he appears disturbed by awareness of failure. Particularly in the case of subjects who are tested by reason of authority³ rather than of their own volition, the examiner is alert to small changes in facial expression or posture, tonal inflection, and other signs of emotional disturbance. Older subjects are not only better able to judge the quality of their own performance than younger ones are; they are more conscious of the possible bearing of the results of the test upon the treatment they may receive later or upon their own prestige in the eyes of the examiner. They are also more adept at concealing their feelings.

A few differences in the conduct of an examination of older subjects as compared to that just outlined for young children may be noted. The side-to-side position of examiner and subject is no longer preferred

³ Such as that of the school or court, parents or guardians, or others responsible for the subject.

except in special cases. Many examiners now place the subject opposite them, but others find that a right-angle arrangement in which the subject sits at the end of a table with the examiner at the side is less embarrassing to those who are nervous or self-conscious.

Supplementary equipment will be suited to the age of the subjects. A side table may contain a few books and magazines of current interest, a curio or two, or, if the subjects are for the most part children of school age, one or two mechanical puzzles or other objects likely to arouse their interest. While these are not used during the test itself, they add to the attractiveness of the room and may help to establish friendly relations with nervous or hostile subjects.

Exact adherence to the test instructions is essential no matter what the age of the subject may be. These instructions should be memorized to a point where no more than an occasional quick glance at the manual is necessary. It is difficult, if not impossible, to establish a complete feeling of ease on the part of a subject who is reminded of the fact that he is being tested by seeing the examiner continually resort to printed instructions for his procedures. Much of the potential advantage of the individual test over the group test consists in its apparent informality, which is lost when the examiner reads his questions instead of asking them more directly. Moreover, too great dependence upon printed instructions necessarily diverts much of the examiner's time and attention from observing the reactions of the subject and thereby does away with the second point of superiority of the individual examination as compared to the group test. The more completely the examiner has reduced the mechanics of the examination to a matter of routine, the more closely his attention can be centered upon the subject.

Other points mentioned in the previous section which have equal bearing upon the examination of older subjects may be noted. A low-pitched voice, a pleasant, friendly manner, adequate but not effusive praise, the establishment of a mental set for each new type of question, courteous replies to the subject's questions, and acknowledgment of his responses to the tests with care to avoid giving cues that may affect his performance on those to follow are equally important in the testing of older and younger subjects. Much that has been said so far in this chapter may be summed up in a few sentences. The good examiner is meticulous about details because he realizes their importance in the total situation with which he has to deal. The poor examiner fails to see this relationship. He may be careless, trusting to the inspiration of the moment to compensate for lack of preparation that should have been made earlier, or he may become so involved in small details that he loses sight of the reason for them. The essence of a good test is the maintenance

of a proper balance between clinical insight and the techniques which make such insight possible.

NOTES ON GROUP TESTING

The greater number of group tests are so planned that they can be administered by persons with only a small amount of training. The apparent simplicity of the procedures has led many to feel that *no* special preparation for giving them is required, and that the question of rapport which is recognized to be a matter of major importance in individual testing vanishes like the rabbit in the prestidigitator's hat when the subjects are to be tested in groups. That this is not the case has been amply demonstrated by Hurlock (1925), who found marked changes as a result of either praise or reproof in the retest performances of elementary school children on a group intelligence test.

In group testing as well as in individual testing, the examiner should be keenly alive to the physical and emotional condition of his subjects. Children, particularly those in the lower school grades, should be allowed to satisfy their bodily needs—to go to the toilet or to get a drink—before starting the test. Desks should be cleared and any necessary adjustments in room lighting or temperature should also be made. All these details should be attended to in an easy, natural manner, not with an air of preparing for a major operation.

Modern children are in most cases sufficiently accustomed to taking group tests to render only a few words of explanation or preparation necessary before beginning the formal instructions. These should always have been carefully practiced in advance before a critical audience of at least one person. While the instructions may be read without their having been memorized, the examiner should be so familiar with the wording that hesitation or stumbling will never occur and that the right words will be emphasized and pauses made at such points as will lend utmost clarity to the meaning. No one, not even the best reader, can guarantee that this will be done unless he gives some preliminary consideration to the matter and devotes enough time to practicing the particular sentences to be read to ensure a smooth and even performance. Particular attention should be given to reading the examples, which usually form part of the instructions for each subtest, in such a way as to make sure that the essential point of each will be clearly brought out.

While the test is in progress, the examiner should be on the alert to note any instances of unusual behavior on the part of the subjects. In the case of a timed test, it may be inadvisable to attempt to deal directly with an individual child who attempts to copy from his neighbor's paper

except as it may be possible to catch his eye and show by a glance that his actions have been observed. Note should always be made of such cases. By a later examination of the two papers for identical errors, an estimate can usually be made of the extent to which the score in question was affected by cheating. Other types of behavior, such as evidences of emotional upset on the part of certain children, should also be noted.

Scoring, including the transmutation of raw scores into interpretative units such as mental ages or standard scores, should be carefully rechecked, if possible, by a different person. Children whose test scores are found to depart markedly from that which might be expected on the basis of the quality of their schoolwork, or such other evidences of ability as may be available, should be retested. If a properly qualified psychologist is available, an individual examination should be arranged for; if this is not feasible, another form of the same group test or a group test of another kind may be tried. If a discrepancy still exists, an unbiased attempt should be made to ascertain its basis. No test is infallible; no human judgment is free from the possibility of error. If the facts are carefully examined with the above points in mind, a satisfactory explanation for most difficulties of this kind can be reached and an estimate of the child's true standing can be made that will usually approach fairly close to the facts.

Intelligence Tests

It is unnecessary to present more than a general outline of the kind of tests used at the present time for the measurement of "general intelligence." Enough has been said in previous chapters to show how the concept of general intelligence was gradually built up and how, in spite of disagreements as to its nature and organization, a system of useful devices for the appraisal of its level of development in the individual has been constructed. In this chapter we shall make a very brief survey of the kinds of tests suitable for use at various age levels and indicate some of the special problems associated with each.

TESTS FOR INFANTS

As here used, the term "infant" applies to those children for whom speech has not yet become a useful means of social intercourse. Generally speaking, this will include babies under the age of about eighteen months, but some infants remain in the prelinguistic stage for a longer period than this, while a few develop enough facility in the use of speech to make it possible to use tests requiring some linguistic skill by the age of fifteen or sixteen months.

Unquestionably, the person who has made the chief contribution to our knowledge of mental development during infancy is Arnold Gesell of Yale University. Although Gesell's tests of mental development lack statistical refinement and his instructions for giving and scoring them leave so much to individual judgment as to render it highly unlikely that different examiners will obtain similar results for the same cases, his work nevertheless has been the chief source of ideas for others whose technical skill in test construction outruns their talent for observing the infinite minutiae of infantile behavior and for judging which aspects of the child's responses to specific stimuli are most significant as indicators of his mental progress.

It is not an easy matter to give a mental test to a young infant. The task is even more difficult when the child has reached an age at which he is aware of strangers and strange places and when the nature of the tasks



Above DAVID WECHSLER
The Wechsler-Bellevue
Scales for adolescents and
adults.



Above RACHEL STUTSMAN
BALL
The Merrill-Palmer Tests
for children of preschool
age.



Below GRACE ARTHUR
The Arthur Point Perform-
ance Scales for children in
the elementary school.

MAUD MERRILL

Co-author with L. M. Terman of the
1937 Revision of the Stanford-Binet
Scales for ages two years and upward.

Below ARNOLD GESELL
Tests for infants.

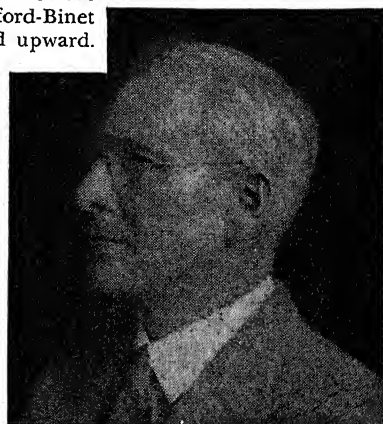


FIG. 20. SOME OF THE AUTHORS OF MODERN INTELLIGENCE TESTS.

requires some kind of active response on his part. For the infant is in no way interested in his own performance as such. Success or failure means nothing to him. Unless the stimulus presented arouses his immediate interest, he will have none of it.

Some of the reasons for the generally low predictive value of the tests used for infants have been pointed out in previous chapters. We may recapitulate these, together with certain others, as follows: (1) the impossibility of guiding the child's response by means of verbal instructions; (2) uncertain and variable motivation; (3) difficulty of controlling and directing the child's attention; (4) observational errors in scoring; (5) lack of an immediately useful criterion for selecting and weighing test items;¹ (6) statistical problems arising from the low chronological age of the children which makes small age difference count for much;² and (7) the unsettled question as to whether or not true intelligence may be said to have emerged before the symbolic processes exemplified in speech have become established. Attempting to measure infantile intelligence may be like trying to measure a boy's beard at the age of three. There is no question but that marked individual differences exist in the rate at which babies acquire the various skills that are appropriate to infancy, that they differ in responsiveness to external stimuli and in apparent alertness. But it will be recalled that equally marked individual differences were found among school children with respect to the sensorimotor tests so hopefully tried out by American psychologists around the turn of the century. These differences, however, were not found to be associated with differences in schoolroom accomplishments or with other evidences of mental ability.

Among the tests for infants at present available may be mentioned the Bühler Baby-Tests (1932), the California First Year Mental Scale (1933) by Bayley, and Psyche Cattell's attempt to extend the 1937 Stanford-Binet to the infant level (1940). None of these has shown an appreciable relationship to later mental standing for children tested before the age of twelve months. For tests given between the ages of twelve and eighteen months, the correlations with the Stanford-Binet

¹ The only really sound criterion appears to be the child's standing on intelligence tests of recognized value after he has reached an age for which the latter tests are appropriate. Although a good many studies have been made to ascertain the extent to which tests made in infancy predict later standing, no one, as far as I am aware, has actually used the predictive value of the separate items as a criterion for item selection when standardizing a new scale.

² In this connection the differences between age reckoned from the date of conception and age reckoned from date of birth should be considered. The conceptional ages of two children born on the same day may differ by a month or more, even though neither is classed as having been prematurely born.

tests given to the same children at three years or older have generally been significantly positive but low, ranging from about $+.35$ to $+.65$.

A few examples of the items used at the different ages follow.

<i>Age</i>	<i>Response</i>
2 months	Eyes follow moving person.
4 months	Lifts head and chest by arms when in prone position.
8 months	Picks up small sugar pellet from table. (Any method of securing it is credited as success.)
12 months	Beats two spoons together in imitation of examiner.
16 months	Puts wooden beads in box.

All the items listed above are taken from the Cattell series. Explicit directions for administration and scoring are given in Cattell's manual.

TESTS FOR CHILDREN OF PRESCHOOL AGE

The chronological age period for which the tests discussed here are suitable extends from about eighteen months to five years. These limits, however, are not absolute but vary to some extent with individual ability.

Many psychologists have designated the ages from two to four years as the "negative phase" in child development. As a general term this would seem to be a misnomer, but it is unquestionably true that it is a time when many children are very hard to handle, especially when their cooperation in a more or less rigidly defined task is to be gained. In spite of every effort to make the materials and the tasks as alluring as possible, a negativistic attitude on the part of the child constitutes one of the most difficult problems for the examiner of young children. The suggestions offered in the preceding chapter for handling children in the test situation were intended particularly for examiners who have not had much contact with subjects of tender years, but here, as elsewhere, there is no substitute for actual experience.

Not only shyness and negativism but the young child's relatively undeveloped awareness of success and failure and his lack of interest in solving problems for the sake of mastery complicate the test situation at these early ages. Motivation is almost wholly a matter of the intrinsic attractiveness of the tasks set, and a desire—not always to be depended on—to please an examiner who has won the child's confidence and esteem. Both are variable factors, for a task which appeals to one child may not attract another, and the relationship between examiner and child may be affected by a number of factors, not all of which can always be controlled. There can be no doubt whatever that a considerable portion of the instability of test scores earned at these ages is attributable to variable interest and effort on the part of the child at successive testings.

Modern tests for preschool children may be divided into several different classes: (1) those of the Binet type, among which the 1937 Stanford, which includes six test items at each half-year period from two to five years, holds first place. The Cattell tests mentioned in the preceding section are more closely calibrated, with six items at each two-month interval between the ages of twelve and twenty-four months, and at three-month intervals to thirty months. The Kuhlmann 1939 Revision is less widely used; (2) tests involving very little use of language, such as the Merrill-Palmer series of performance tests (1930); (3) tests involving both verbal and nonverbal responses which are scored separately, such as the Minnesota Preschool Scales (Goodenough, Maurer, and Van Wagenen, 1932, revised edition, 1940); and (4) wholly nonverbal tests, such as the Atkins Object Fitting Test (1931). Hildreth (1939, 1945) lists 33 different tests of intelligence for the preschool ages, as well as many others for the measurement of related abilities, such as speech tests, block-building tests, and so on.

Most studies have shown that tests given between the ages of eighteen months and five years have some predictive value for later mental ability and that the magnitude of the correlations obtained between early and later tests increases regularly with advancing age and with the shortness of the interval between testings. Goodenough and Maurer (1942) found that the nonverbal scales of the Minnesota Preschool Scales predicted later mental standing more accurately than did the verbal scales. This is not in accordance with the findings for older children and may very possibly be a result of the greater interest value for young children of tests making use of concrete material. By far the most important study of IQ constancy over a considerable period of time, however, is that by Bradway (1944), who, ten years after the first testing, re-examined as many as could be located of the children of preschool age used in the original standardization of the 1937 Stanford-Binet. She obtained correlations averaging slightly over $+.60$ for different combinations of Forms L and M at various age levels. The fact that about one child in four showed an IQ change as large as 15 points led Bradway to advise that "an individual IQ obtained before the age of six must be interpreted with discretion." This is assuredly wise counsel.

TESTS FOR KINDERGARTEN CHILDREN

It has not been found possible to develop tests for children under the age of five which can be given to more than one child at a time. After a few months of kindergarten experience, however, small groups of children may be tested simultaneously, and a number of tests have



FIG. 21. SOME OF THE MATERIALS USED IN THE MINNESOTA PRESCHOOL SCALES. (Courtesy of the Educational Test Bureau, Minneapolis.)

been devised for which the procedure is simple enough for the tests to be given by the classroom teacher. Two which utilize somewhat different procedures from most others are the Goodenough "Draw a Man" test (1926) and the Thurstone Primary Abilities Tests for Ages Five and Six, of which mention was made in a previous chapter. The Stanford-Binet, however, remains the standard for this age. Of the nonverbal tests, the Merrill-Palmer is perhaps the most satisfactory for children whose ability is below average; for those of average or superior ability, the Arthur Point Performance Scale may be used.

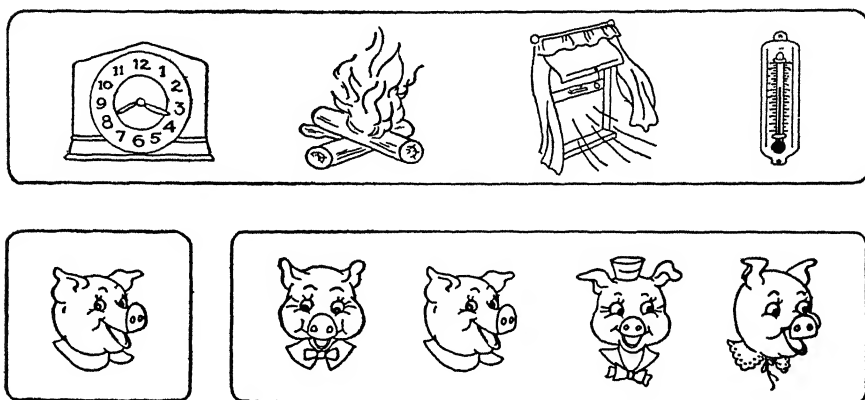


FIG. 22. TWO OF THE ITEMS FROM THE TEST OF PRIMARY MENTAL ABILITIES FOR AGES FIVE AND SIX BY THELMA GWYNN AND L. L. THURSTONE. The first row of pictures is from the test of word-meaning. The instructions are: "Mark the picture that answers this question: Which one do you look at if you want to know how cold it is? Mark it."

The second row is from the test of perceptual speed. The instructions are: "In every row of pictures you are to do two things. First mark the picture all by itself in the little box. Then find the picture in the big box which is exactly like the picture in the little box and mark it too. Work fast. Do as many as you can on these two pages before I tell you to stop. Are you ready? Begin! (Reprinted by permission of the authors and of Science Research Associates, the publishers.)

TESTS FOR ELEMENTARY SCHOOL CHILDREN

In many ways the elementary school period, especially the ages from eight to fourteen, is the optimal time for measuring the degree of brightness. By this time, children have advanced to a stage at which such personality factors as shyness and negativism rarely create special difficulties for the examiner, and their interest in the tests as *tests* of their own knowledge and skill is keen. They have a strong sense of competition, even with respect to unseen and unknown adversaries, and are intensely interested in making a high score with only moderate regard for the

nature of the problem. Exceptions occur, of course. There are plenty of elementary school children who adopt an indifferent or rebellious attitude toward tests and tester, but on the average this is the time when strong motivation and good cooperation are easiest to secure. As compared with older children or adults, children in the elementary school are less likely to be suspicious of the purpose of the test, to wonder "what it's all about," or to be anxious and self-conscious about their own performance. They want to do well but are more ready to take failures in their stride.

Not only are children at this age relatively easy to test but the tests available for the purpose of measuring their intellectual ability have certain advantages over those designed for the earlier or the later years. Such tests as the Stanford-Binet are at their best over these age ranges. Test makers have had the great advantage of a longer period of experimentation and trial, with reasonably satisfactory criteria against which results could be checked. Being near the middle of the age range covered, scores are less likely to be affected by too close approach to the floor or the ceiling of the scales. And while interests and abilities show a considerable degree of specialization even at these ages, such specialization has not yet crystallized into differentiated courses of study or into the different occupations of adult life. General intelligence is still manifested in a more nearly uniform way.

The Stanford 1937 Revision remains the most dependable of the available measures for these age ranges. As a matter of fact, if only the range from six to fourteen is considered, and if correction is made for the inequalities in the standard deviations that were mentioned in a previous chapter, this test approaches very closely to the requirements for a standard instrument. Other revisions of the Binet, such as the Kuhlmann 1939 Revision or the now little-used Herring Revision (1922), have never been serious rivals. Objection to the Kuhlmann Revision lies first of all in its emphasis upon speed of response, which makes timing in terms of seconds necessary for many of the items. When only an ordinary stop watch is used, errors of timing are likely to occur, and for many of the tests an error of a single second has an appreciable effect upon the score. The method of scoring is complicated, demanding many divisions, additions, and other arithmetical processes as well as continued reference to tables of standards, all of which is conducive to mechanical errors. The scale as a whole is heavily weighted with formalized items, such as crossing out specified letters in a pied text, counting dots, recognition of geometrical designs, and so on. Although Kuhlmann expressed a definite preference for the Heinis PC, the test may also be scored in terms of mental-age units from which IQ's may be computed. Little

information is available with regard to the stability of its results or to its predictive value in estimating later mental status.

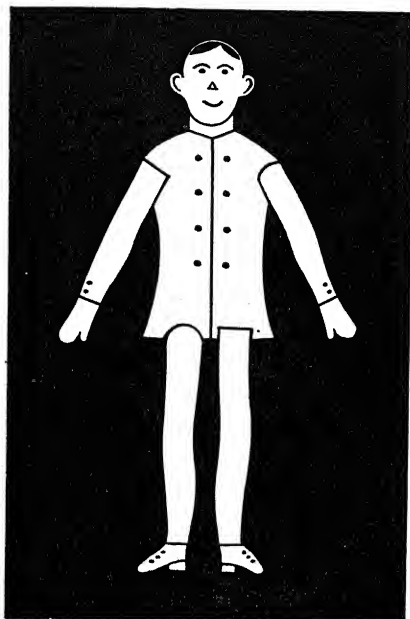
Many group tests for these ages are also available. Of these the Kuhlmann-Anderson, now in its fifth edition (1942), is probably the most popular. The distinctive feature of this test is its use of overlapping batteries for the successive grades. For each grade from I to IX there are ten subtests with varying numbers of items in each. Some of the subtests at each year level duplicate those used for the previous year, while others are introduced for the first time. In some of the earlier editions, scoring was in terms of the Heinis PC; but the procedure now favored is the median mental-age method from which IQ's are computed in the ordinary way, although data are wanting to show whether or not the conditions necessary for the valid use of the IQ have been fulfilled.

Of the nonverbal tests for this period, the Arthur Point Performance Scale is most commonly used. For ages beyond nine or ten years, however, both the self-correlations and the correlations with other criteria such as the Stanford or the Kuhlmann-Binet are low (see Arthur, 1933). Among the group tests not requiring the use of language, the Good-enough "Draw a Man" test is much used for the first two grades and the Pintner Nonlanguage Mental Tests (1919) for the upper grades.³

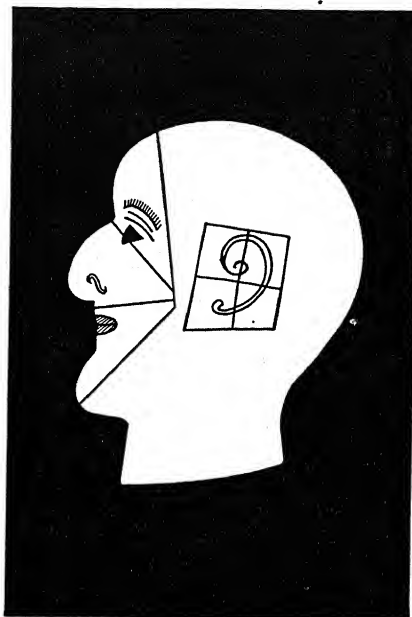
TESTS FOR HIGH SCHOOL AND COLLEGE STUDENTS

Although the 1937 Stanford-Binet includes tests for age fourteen, for the "average adult," and three levels of difficulty for the "superior adult," this test is better adapted for use with children than for older adolescents or adults. For use beyond the junior high school, the Wechsler-Bellevue, originally published in 1939 and since revised (1944), is steadily gaining in popularity. Like the Stanford, the Wechsler-Bellevue is administered individually and results are expressed in terms of "IQ." Wechsler, however, has used this term only in order to take advantage of its widespread popularity. As used by him it is a derivative of the standard score similar to the IQ Equivalent used in the Minnesota Preschool Scales. Tables of standards for transmuting into "IQ's" the number of items correctly passed are presented in the *Manual*. Although these standards cover the ages from six to sixty years, the test is little used with children below the high school age.

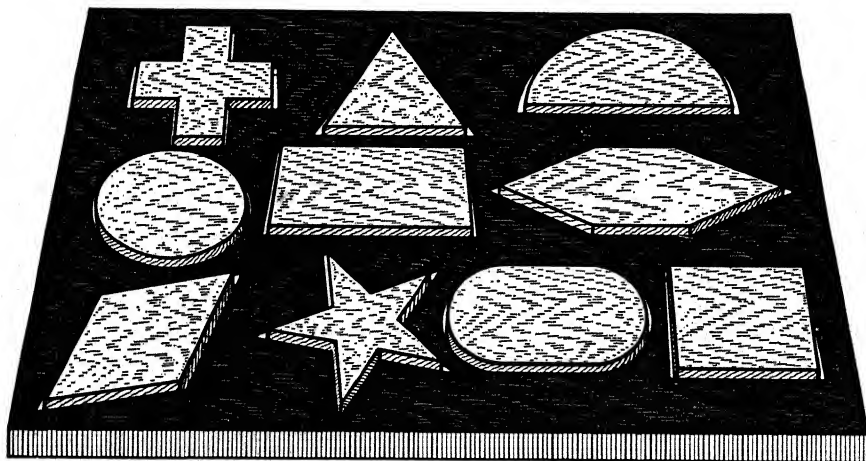
³ Neither of the above tests can be depended on to give more than a very rough estimate of the mental level of the subjects tested. A good nonlanguage group test for the elementary school ranges would constitute a real contribution to the equipment of the mental examiner.



THE MANIKIN TEST



THE FEATURE PROFILE TEST



THE SEGUIN FORM BOARD

FIG. 23. SOME OF THE TESTS COMPRISED IN THE ARTHUR POINT PERFORMANCE SCALE.
(After photographs provided by the C. H. Stoelting Company.)

The Wechsler scale is made up of eleven subtests, of which six constitute a verbal scale and the remaining five a nonverbal scale. The verbal subtests include a test of general information, a comprehension test, an arithmetic test, a digits test, a vocabulary test, and a test of similarities. The nonverbal scale includes a test of picture completion and one of picture arrangement, an object assembly test, a block design test, and a digit-symbol test. Separate norms for each of the two scales and for the total or "full scale" are available.

A distinctive feature of this test is the use of a kind of pattern analysis made possible by equating all scores to a common basis in terms of their deviation from the mean of the age group to which the subject belongs. A study of the differential patterns of response made by persons with certain known characteristics has resulted in the elaboration of a system of "signs" believed to be diagnostic of various personality trends or mental abnormalities. These will be mentioned further in a later chapter.

The Wechsler nonverbal scale provides what is probably the best answer to the difficult question of finding a useful measure of intelligence for persons with a limited knowledge of English, for the deaf, or for others likely to be unfairly handicapped by a scale requiring the use of language. Most of the so-called "performance tests" have little value as indicators of general intelligence after a certain limit of proficiency has been reached. Although there is evidence that both the self-correlations and the correlations with other criteria of the Wechsler nonverbal scale are somewhat lower than those for the verbal scale, it still appears to be superior to most other nonlanguage tests available for the older age range.

Because of restrictions imposed by time and funds, group tests are widely used for purposes of educational and vocational guidance at the high school level. Among the most popular of those now used are the Terman-McNemar Test of Mental Ability (1941), which is a revision of the widely used Terman Group Test that appeared in 1920; the Ohio State University Psychological Test, which has been revised in each alternate year since 1919 and is widely used as a basis for predicting the likelihood of college success; the Wells Revision of the Army Alpha (1932); and the college entrance examinations prepared and revised annually by the American Council on Education. Many colleges and universities have prepared their own entrance examinations. The I.E.R. intelligence tests prepared by the Institute of Educational Research of Teachers College, Columbia, under the direction of E. L. Thorndike (1935) are also popular.

Both at the high school and at the college level, however, tests of special aptitudes and interests and tests designed to throw light upon

personal-social relationships and adjustment are likely to be more informative than the "global" types of intelligence test for the majority of students. Not that general intelligence is unimportant for college success. Far from it! But by the time of college entrance the steady process of screening that has been going on since the primary grades has gradually weeded out most of the subjects whose general intellectual level is much below the top quartile of the population at large. The great bulk of the college population is made up of the upper 10 per cent of the general population. Entrance requirements in practically all universities of recognized standing are high enough to exclude most really backward students, although it is true that an occasional person will slip in by reason of family pressure, political influence, athletic prowess, or other reasons.⁴ When other adverse factors, such as the necessity for partial self-support while attending college, lack of interest in college work, or serious emotional disturbances, are present, students whose intellectual level is near the lower boundary of their college group may fail when those of higher general ability would succeed under the same conditions. General intelligence tests thus have a real place in the college testing service, even though it is a place of smaller importance than that assigned to them at an earlier stage of educational progress.

TESTS FOR UNSELECTED ADULTS

The outstanding examples of a nation-wide testing program for adults are to be found in the psychological examinations of American soldiers during World Wars I and II. These tests will be discussed in Chapter 33. Apart from military and industrial uses, or occasional instances where the mental level of a subject who has been convicted of a criminal offense is questioned, general intelligence tests play but a small part in the lives of most adults whose formal schooling has been completed. When used, the selection of tests will depend for the most part on the educational level of the persons to be tested and on the amount of time and funds available for testing. When resources permit the use of an individual test, the Wechsler-Bellevue is probably best adapted to this purpose since it was standardized on adults whose educational background varied widely, and is scored for a number of special aspects of ability as well as for general mental level. When means are more limited, one or another of the leading group tests is likely to be employed. The

⁴ An occasional instance has been reported in which students whose mentality was so low that many would regard them as genuinely feeble-minded have succeeded in entering college. As far as I know, none of these students have proved able to pass any of their academic courses.

Wells Revision of the Army Alpha and the Otis Self-Administering Test (1922) are among those most often used.

MENTAL TESTS IN SENESENCE AND OLD AGE

Among the questions that have interested students of human behavior for many years is that of the rate and manner of mental decline with advancing age. The point is one that has a bearing upon many social and political practices; it affects family life and organization as well as the national economy. The scientific importance of the problem is unquestionably very great.

Dependable facts, however, are not easy to secure. For truly sound conclusions the same persons should be examined both before and at varying ages after mental decline has begun. Just where the turning point lies is not easy to ascertain. The only sound facts that have emerged from the many studies made seem to be these: (1) For most aspects of ability the growth curve reaches a peak in fairly early maturity, then undergoes a well-nigh imperceptible decline for some years during which, for most practical purposes, the level may be regarded as stationary. Thereafter the descent gains force as age advances. (2) Certain aspects of ability decline earlier and at a much more rapid rate than others. (3) Some persons maintain their early mental vigor for years after the majority are well advanced on the downward slope. All these factors make it difficult to establish general rules as to the age at which mental decline sets in or the rate at which it advances. The results of specific investigations will therefore vary according to the content of the tests used, the extent to which the results are modified by such factors as speed of reading, which is likely to be adversely affected by the visual defects common among the aged, the amount of writing required, the familiarity of the material, and other matters not closely related to general intellectual ability.⁵

The only test that takes definite account of the qualitative as well as the quantitative aspects of mental decline in the later ages is the Wechsler-Bellevue, which presents different standards for the ages from twenty to sixty as well as for the younger ages. In *The measurement of adult intelligence* (Revised edition, 1944) Wechsler provides a good deal of valuable information on the nature of the age changes that occur with advancing years.

The use of different standards for adults of different ages raises a question of considerable practical as well as theoretical importance.

⁵ For a brief review of the chief findings in this area see Chapters 22 and 23 in *Developmental psychology* by Goodenough (1945).

Suppose that Individual A, aged twenty-five, and Individual B, aged sixty, are found to have identical "IQ's" according to the Wechsler-Bellevue standards. Their actual test performances, however, differ considerably, both quantitatively and qualitatively. How are such results to be interpreted? The use of different age standards in childhood poses no such question, for children compete only with those who are near their own age. But this is not the case among mature individuals. Just what may be the best method of handling the changes that occur during later life in the appraisal of individual mental ability is a question that awaits further study.

Tests of Educational Aptitude and Achievement

"READINESS" TESTS

During the decade of the 1920's much interest in the relation of school achievement to standing on intelligence tests was displayed by workers in the field of educational research. The high correlation usually found between the two led quite naturally to the belief that intelligence tests may provide information of much value for school organization and for adapting the methods of instruction to the differing capacities of the children. This gave rise to various plans of ability grouping and of individualized instruction which aroused much interest among the more progressive educators.

One of the most disturbing educational problems of that day was the large percentage of failures in the first grade. In some school systems it was found that from 25 to 50 per cent of the children were obliged to repeat the first grade at least once and that there were many who failed twice or even three or four times before reaching a level of accomplishment at which promotion to the second grade was deemed advisable. The discovery that the majority of these children did poorly on intelligence tests as well as in their schoolwork caused many to ask whether there might not be a mental level below which mastery of first-grade work, if not absolutely impossible, was at least so unlikely as to make it unwise to admit children to the first grade until the required level of mental development had been reached. Just what this level might be was not universally agreed upon, but most of those who had examined the figures on school retardation were of the opinion that a mental age of approximately six years was the minimum level at which success in the first grade might reasonably be expected. Since reading is the subject of prime importance in that grade, the question soon resolved itself into the following: What is the lowest mental age at which it is possible for the average child to learn to read under the methods of instruction

commonly used in the public schools?¹ And since a thing may be possible but only at a cost so great as to render it impractical, the question soon became: What is the optimal mental age for a child to receive his first instruction in reading?

Varying answers to both these questions appeared. Davidson (1931) found that, when taught by a special method devised by herself, children with chronological ages of three to five years but all with mental ages of four years were able in each case to make some progress toward learning to read. However, the amount of progress varied considerably from child to child in spite of the uniformity in mental age and in manner of instruction. She also found that the bright three-year-olds learned much more rapidly than did the children of four and five years, and that the four-year-olds, who were of average intelligence for their age, did somewhat better than the retarded five-year-olds. But the variations from child to child were much too great to be explained on the basis of their intellectual level alone. This study and others less well controlled, as well as common observation, led to the belief that while the age and the rate at which a child is able to learn to read are undoubtedly conditioned to a large extent by his intellectual maturity, and although some methods of teaching are likely to be more effective than others, unmeasured differences from child to child in special aptitude for reading still exist. The age at which a child is ready to begin to learn to read is thus determined not only by his general intellectual level but also by the degree of special aptitude he possesses.

Working on this hypothesis, a number of persons attempted to devise scales for the explicit purpose of predicting a child's success in learning to read. The first specific use of the term "reading readiness" seems to be attributable to Deputy (1930), who tried out a series of five tests including both measures of intelligence and certain others thought likely to be more specifically associated with reading aptitude, such as comprehension and recall of stories, word matching, and visual association tests. For 103 children tested at the time of entering first grade, the multiple correlation between test score and a combination of teachers' ratings on quality of schoolwork, together with scores on several primary reading tests at the end of the school year, was found to be $+ .748$, which is rather higher than is generally obtained for intelligence-test scores alone when unselected school children are the subjects.

Other tests of this kind soon followed. By the middle of the 1930's

¹ "Mental age" was usually understood to refer to standing on the 1916 Stanford-Binet. Although a number of group tests for use with first-grade children had appeared by the early 1920's, they were not generally regarded with much favor for individual diagnosis.

some half dozen or more had been published, of which the Metropolitan Readiness Tests by Gertrude Hildreth and N. L. Griffiths (1933) has retained its popularity in spite of a number of other more recent additions to the list, among which that by Gates (1939) is perhaps the most widely used.

As the name suggests, tests of reading readiness are designed to help teachers and school administrators to decide on the advisability of admitting children to the first grade. They are particularly useful in cases such as arise when a child's chronological age is on the border line between that which would ordinarily admit him to the first grade and that which would lead to his retention in kindergarten for another term, or where other special factors leading to doubt as to the wisdom of first-grade entrance are present. These tests are generally planned for administration by kindergarten or first-grade teachers without special training.²

Tests of "readiness" to begin other subjects of the elementary school curriculum have also been worked out. Several "arithmetic readiness" tests have appeared, and there have also been attempts to devise more objective ways of deciding when a child is ready to begin writing, formal spelling, and so on. None of these, however, have proved to be as practically useful as are the tests of reading readiness. A child's success in the first grade is almost wholly determined by his skill in reading. A test which predicts reading accomplishment in advance of trial will therefore be a valuable guide in deciding whether or not he is ready to undertake the work of the first grade. But once he passes into the higher grades, the time at which he begins the various school subjects is determined by the curricular requirements for the successive grades, not by the "readiness" or "unreadiness" of the particular child.³ For this reason, tests of school achievement together with diagnostic tests⁴ designed to uncover the reasons why individual children do poorly in certain subjects are usually preferred for the study of educational progress in the elementary school.

At the high school and college levels, however, where a choice of subject matter becomes possible, a more sophisticated form of predictive

² It is a moot question whether, in such cases, the tests should be given by the teacher who has had charge of the children during the year—in this case, the kindergarten teacher—or by another teacher whose attitude toward them is more impersonal. The classroom teacher will usually be able to secure better rapport, but if her testing procedure is affected by opinions already formed with respect to the relative ability of the children, much of the discriminative potentialities of the test will be lost.

³ For many of the more progressive school systems, where promotion is based chiefly or wholly upon chronological age and a child's interest is deemed a sufficient guide as to what he shall attempt to learn, the above statement is, of course, inapplicable. In these schools, however, the readiness tests have no place since the child's wishes would always take precedence over the test results.

⁴ These tests are discussed in Chapter 23.

testing known as "aptitude measurement" fills an important place. This includes tests of aptitude for a general broad field, for instance, mathematics or foreign languages, as well as more specific tests designed to predict success in limited areas such as plane geometry, shorthand, or chemistry. Some of these tests, for example, tests of clerical or mechanical aptitude, have a vocational as well as an immediately educational reference.

TESTS OF EDUCATIONAL ACHIEVEMENT IN THE ELEMENTARY SCHOOL

Tests of educational achievement differ from aptitude tests in that they are primarily designed to measure the knowledge and skills already acquired. Although prediction of future accomplishment is not infrequently made on the basis of such tests, their fundamental purpose is to ascertain the child's present educational status. Such knowledge provides the teacher with a firm starting point from which she can safely proceed to higher levels. In its absence she is likely to assume much that does not exist, with consequent waste effort both on her own part and on that of the children.

The number of more or less well-standardized school achievement tests for the elementary grades that have appeared during the past three decades will amaze many persons. In the 1939 edition of Hildreth's *Bibliography of mental tests and rating scales*, 138 different tests of arithmetic ability alone are listed, besides those included in general batteries of educational tests. The 1945 *Supplement* adds 23 to the number. Test makers have been equally industrious with respect to the other subject fields. No attempt will be made here to describe these tests in detail. We shall note only some of the forms most commonly used and the special problems which have led to the multiplication of tests and testing methods in areas which at first glance appear relatively simple and straightforward.

Three general ways of measuring educational achievement are in common use. These include (1) *performance scales*, in which the subjects are set a series of tasks, graded in difficulty so as to cover the entire range of ability likely to be found within the group for whom the test is designed; (2) *product scales*, consisting of a standard series of sample products in such fields as drawing, handwriting, sewing, English composition, and the like, which have been scaled in equally spaced units of excellence. When a product scale is used, a specimen of the child's work is compared with each of the standard specimens in turn until a level most nearly corresponding to it in quality is found. The scale value

of that specimen is then taken as the child's score; (3) *information tests* designed to ascertain the extent and precision of a child's knowledge about a topic. These are commonly preferred for the measurement of achievement in such fields as history, geography, and current affairs. Variations and combinations of these methods also occur.

The task of constructing an educational test, however, is by no means solved with the choice of the form in which the material is to be presented. The educational philosophy of the test maker will be mirrored in the tests which he constructs, and unless that philosophy is clearly formulated and well organized, the tests are likely to be similarly lacking in integration and structural design. Here are just a few of the questions which the test maker should ask himself before embarking upon the task of constructing a test, and which the test user should also consider when examining a test which he proposes to try out.

1. Is the field to be measured looked upon as one in which the parts are subordinate to the whole or as one in which the whole is essentially an aggregate of the parts? In the former case the test items will deal largely with consequents and antecedents, with general principles and the interrelationships of facts. In the second case, the emphasis will be upon skills and knowledge and their application to specific and familiar situations. The first is a molar, the second a molecular view of the educational process.

2. How extensive is the universe with which the test is to deal? What are its most salient features? What principles govern the relationship of the parts to each other and to the whole? If the test is intended to be diagnostic, that is, if it is so designed as to show in what specialized aspects of the field measured are to be found the child's points of greatest weakness or strength, the question of the relationship of parts is of prime importance. For if the comprehension of a given fact or principle or the acquisition of a certain skill is essential for the grasp of other principles or skills, a good diagnostic test will give preference to material in that area and if necessary will neglect other topics which may seem equally important in their own right but which are less closely integrated with the remainder of the field.

3. The point just mentioned is closely related to the much-discussed question: How can tests of understanding, as opposed to those dependent chiefly or wholly upon the acquisition of skills and the memorizing of facts, be devised and standardized? The Forty-fifth Yearbook of the National Society for the Study of Education, Part I, on *The measurement of understanding* (1946), includes a number of pregnant discussions on this head, together with practical suggestions for the construction of tests which demand comprehension as well as knowledge. It should

be noted, however, that understanding is not a free-floating mechanism but rests on a body of facts and skills. A child who fails on an item intended to measure comprehension may do so either because he lacks the ability to integrate the knowledge at his command or because he lacks the knowledge essential for understanding the principle. The fact that tests of knowledge which depend chiefly upon memorizing are usually easier to construct has undoubtedly led to the overweighting of most educational tests with factual items to the exclusion of those which demand thought and reasoning, but this does not mean that the former have no place in educational measurement. Particularly in diagnostic tests, both types of approach are needed.

4. Another question to which the answer will vary with different educational theories has to do with the emphasis placed upon speed of response. Although the correlation between speed and accuracy is positive and high, as is also the correlation between the difficulty of the tasks which a subject is able to perform and the rate at which he performs them,⁵ it is nevertheless true that individuals differ in respect to their usual rate of work, irrespective of differences in ability. It is also true that test makers differ in respect to the importance they attach to quickness as such. These differences in attitude toward speed are likely to be reflected in the tests they develop. Some tests have time limits so long that rate of work is of little consequence, since even the slowest workers will usually be able to complete as much of the test as they are able to do. In others, the time set is so short that few or none will reach their performance limit before the signal to stop is given. For some of the drill subjects such as arithmetical computation, speed tests of this kind are unquestionably valuable, but the uncritical assumption that the time limits imposed for a particular test have no bearing on the results obtained by its use is certainly not warranted.

In addition to the immediately practical consideration of time, funds, and personnel which may put some tests out of his reach, the school psychologist or school principal who is attempting to select a

⁵ A good many of the figures reported in the literature on the relationship between "power tests" and "speed tests," that is, between tests for which unlimited time is allowed so that each subject is permitted to complete as much of a test as he is able, and those in which a strict time limit is imposed, are spuriously high since they are based upon the correlation between the amount accomplished within a single short interval and the total completed when unlimited time is permitted. This, of course, involves the correlation of a part with a whole and its magnitude will vary according to the proportion of the whole that is included within the part. It is true that such a correlation has a certain practical significance since it indicates how closely the results of the time-limited procedure approximate those which would have been obtained had unlimited time been allowed. But it tells little about the intrinsic relation between "power" and "speed."

single educational test from among the hundreds competing for his choice will do well to begin by putting his own house in order. He should know by what principles of educational philosophy he is steering his course and how well the tests he uses are adapted to charting his progress in the direction he has chosen. Far too much educational testing is aimless, with both the choice of tests and the uses to which the results are put dictated more by the fashion of the moment, by the fact that "other people are using this and doing that," than by any more reasoned judgment. The selection of tests should be governed not only by a critical examination of the statistical results reported in the literature, but by a thoughtful consideration of the educational philosophies which, wittingly or unwittingly, have determined their form and structure.

EDUCATIONAL TESTS FOR USE IN SECONDARY SCHOOLS AND COLLEGES

Curricula in the elementary school are little, if at all, differentiated. Certain tool subjects are looked upon as basic for all children, regardless of what their later careers may prove to be. Every child, it is believed, should learn the three R's, and should be familiar with some of the elementary facts about the history and geography of his country. Once this fundamental level has been attained, however, the vocational specialization of the future begins to cast its shadow upon the pattern of education of the children who must prepare for it. As early as the junior high school, in some cities, vocational courses begin to occupy a considerable part of the time of certain children who, it is already apparent, are unlikely to attend college and whose interest in the more abstract subjects of the school curriculum is small. By the time the senior high school is reached, college preparatory courses are rather sharply distinguished from those not designed to prepare for college admission. Within the former group, the curricula of students planning to major in one of the sciences will differ in some respects from that of students looking forward to entering a liberal arts college. Among the high school curricula not designed to prepare students for college are distinguished the business courses planned to fit students for minor office positions immediately upon graduation or to prepare them for entrance to business schools giving more advanced training, and courses filling a similar role with regard to the mechanical trades. Still further opportunities for specialization of training are offered by some of the large city high schools.

When the choice of high school courses is left entirely to the student and his parents, many unwise selections will inevitably occur. The choice

of a preparatory curriculum will be governed by all sorts of factors having no relationship either to the general ability and special interests and aptitudes of the student or to the special circumstances connected with the vocation selected. At no other point in the child's educational career is the need for educational and vocational counseling so urgent, for the choice of a high school curriculum may well prove to be a decisive factor in his entire future life.

Educational counseling in the high school is a task demanding thorough technical training, warm personal relations with students and parents, and an ability to see the child in relation to his social setting and not merely as a conglomeration of test scores and ratings. Not that the scores should be neglected. Here, if ever, a thoroughgoing examination of the child's potentialities and accomplishments in terms not only of general intellectual ability but of special interests and aptitudes, the knowledge and skills which he has already attained and the weaknesses and defects which, if uncorrected, would hamper further progress, as well as the personal-social characteristics which have so important a bearing upon his chances of success, is called for. A program of this kind is costly, but economy is not likely to be achieved by employing counselors of a lower degree of competence merely because they can be had more cheaply. A poor choice of tests to be used, with unprofitable overlapping of results in certain areas and neglect of others equally important, inefficient methods of scoring and recording results, and other wasteful procedures which a better-trained person could have avoided will soon eat up the difference in cost, while the loss resulting from misleading or inadequate appraisal of the abilities of the students is irreparable.

A testing program for high school students should include:

1. A dependable measure of general intellectual ability.
2. Tests of special vocational interests and aptitudes.
3. A battery of tests to measure proficiency in the tool subjects of the elementary school.
4. When necessary, diagnostic tests to uncover the nature and sources of special weaknesses in the tool subjects.
5. Tests of personal-social characteristics selected with special reference to vocational choices.
6. Information regarding family history and home background, health record, measurements of height and weight, notes on any special physical defects, and a series of ratings on behavior and personal characteristics by at least two teachers.
7. Records of all high school grades.

8. A well-planned record system from which any known fact concerning a particular student may be quickly obtained.

Ideally, a high school or college testing bureau should provide psychological service for the entire student body. In practice, this ideal is seldom accomplished. As a rule, only those students who voluntarily seek its aid, and those who are referred by members of the faculty because of some peculiarity of conduct or unexplained failures in their course work, will have more than formal association with the bureau. Most colleges give some kind of college ability test to entering freshmen, on the basis of which students who earn very low scores may be refused admission, or at least strongly advised against entering, or they may be placed on probation for a term. College practices, however, differ widely in such matters. Some colleges use their own entrance tests; others make use of the College Board Examinations, which have been devised and standardized by a group of experts and are undoubtedly among the best of their kind.

Standard tests for measuring proficiency in most of the leading college and high school subjects are available and some teachers find these tests useful. However, differences in the content of textbooks used and in the manner and order of presenting the material introduce a good many difficulties since there is much greater variability among college courses called by the same name than is to be found in elementary schools and high schools. Many teachers prefer to devise their own examinations. For this purpose the so-called "objective" form of presenting the questions has gained steadily in popularity.

The objective examination differs from those commonly used a few decades ago in that the student is not required to formulate the answers to questions himself. He need only make a selection from among several alternative answers or express a judgment (true or false, right or wrong) concerning certain statements presented for his criticism. At most, he may be asked to supply an omitted word in a sentence or a paragraph, or in some cases to solve problems dealing with the course material. But he will not be required to organize his knowledge in the form of a connected discussion of a topic, to describe phenomena without the aid of formal cues, or to offer more than a fragmentary comment or criticism in predesignated terminology on any question or theory.

The objective examination has many advantages. When class enrollment runs up into the hundreds, the task of grading, in a reasonably fair and meaningful way, a set of papers, averaging ten or twelve pages of more or less illegible handwritten answers, becomes well-nigh impossible. Examiners are human, and as fatigue and perhaps boredom

set in, their standards, unless rigidly fixed by some external criterion, are almost certain to change. Some of the answers are likely to be ambiguous; in others the real point will be so concealed in a mass of verbiage that it may easily be overlooked. Some students are adept at covering up their ignorance by talking around a subject; they confuse the issue by a display of related knowledge but do not answer the question that was put them. When college classes were smaller, the alert teacher was in a better position to detect and handle such cases, but as enrollment has increased, the amount of time that it is possible to spend on the grading of a single paper has decreased proportionately with a resultant decrement in accuracy of grading.

All this has led to the substitution of formalized tests modeled after the same pattern as the educational tests and scales handled by commercial publishers in place of the "essay-type" tests of an earlier generation. Although some teachers still prefer the latter type and some a combination of the two, the use of the objective examination appears to be steadily on the increase, and a rather extensive body of literature dealing with the construction and validation of tests of this kind has accumulated.

The relative merits of the two types of approach to the appraisal of student achievement is a matter of controversy. The objective type may become a necessary expedient under the conditions now prevailing in many of the larger colleges and universities. The scoring can be done by clerks or even by machines. Since little or no writing is required from the students, who need only check or underline their choice of responses, the number of questions that can be answered within the allotted time period is many times as great as is possible when the essay type of examination is used. By comparison, however, these questions are usually brief and fragmentary.

The major criticisms of the objective examination are these: (1) It frequently leads students to less desirable forms of study habits, to a memorizing of facts without much regard for their context. (2) Since no opportunity is provided, as a rule, for shades of meaning to be expressed or for modifying circumstances to be indicated, it is likely to stimulate uncritical and somewhat dogmatic ways of thinking. (3) It provides the instructor with little or no basis for judging the extent to which students have organized their knowledge into major units or have grasped more than very simple and elementary relationships and meanings. (4) In many cases the instructors who prepare these examinations are ignorant of even the elementary rules for formulating questions of this kind. The result is that many of the questions are likely to be ambiguous, in some the right answer can easily be found by eliminating those alternatives that are obviously wrong, or the wording of one question will provide

the answer to another. In their search for complete objectivity, some test constructors lose sight of the dictates of common sense. For example, in their use of completion tests, some of these enthusiasts indicate by dots the number of letters in the word to be supplied.⁶ The student who chances to think of a synonymous expression is out of luck, even though it may actually be preferable to the one which the examiner had chosen.

None of these criticisms necessarily holds good for all such tests, but in too many instances they are amply justified. The preparation of a good objective examination calls for technical knowledge which the average college instructor lacks. The ease with which such papers can be scored is tempting to the busy teacher who dislikes the recurrent chore of reading the long examination papers formerly demanded. But he is likely to be dismayed by the idea of spending an equal or greater amount of time in order to plan and construct an examination of the objective type which will conform to the basic principles and rules that must be met if satisfactory results are to be obtained.

Criticisms of the essay type of examination center for the most part about the difficulty of devising and maintaining a scoring procedure that will be equally fair for all students in spite of the great diversity in the form and content of their answers. A number of people have maintained that by drawing up a standard code of required points against which the replies are to be checked, most of the advantages of the objective type of examination may be had. Except for the time factor when large numbers of papers are to be read and scored, there is evidence that this may be done, and unless the examiner is willing to make the necessary effort to learn how to construct the objective type of examination and to make practical use of the knowledge, the procedure just mentioned may be preferable.⁷

Reading makes up a large part of the requirements in almost all college and university courses and the amount of required and recommended readings is so great that the student who reads slowly is at a

⁶ In one such case with which I am personally acquainted, the students had prepared lists of words likely to comprise the missing terms. These lists were arranged according to the number of letters in the words—five-letter words, six-letter words, and so on, and were handed down from generation to generation of students, who conscientiously memorized them by way of preparation for examinations.

⁷ One of the earliest comprehensive monographs on the construction of objective examinations is that by Paterson (1925). More recent discussions of the same subject have been presented by Hawkes, Lindquist, and Mann (1936) and by Ross (1944). A valuable series of articles dealing with methods of overcoming many of the objections to newer types of examination, particularly those having to do with its limitation to factual types of questions is to be found in Part I of the Forty-fifth Yearbook of the National Society for the Study of Education entitled *The measurement of understanding* (1946). Many other references can be located by examining the files of *Psychological Abstracts*.

considerable disadvantage when compared to his mates. Differences in reading rates among college students are much greater than most people realize, nor is the popular idea that slow reading makes for better understanding warranted by the facts. Tests for measuring the rate of silent reading and the level of comprehension of that which is read are supplemented by photographic methods of measuring eye-movements during reading, devices for measuring the span of visual apprehension, and other mechanical aids for studying the problem. When slow reading is due chiefly to bad reading habits in persons of good general ability, much can be done by systematic efforts toward increasing speed, but in many cases the difficulty is more deep-seated. Sluggish thinking and poor comprehension are not easy to correct by training.

THE COMPARISON OF ABILITY WITH ACCOMPLISHMENT

That children of inferior intelligence are slower to acquire the educational skills taught in the classroom than are their better-endowed mates has long been known. Teachers recognize this, at least in principle, and are ready to agree that required standards should vary with ability. But how to equate the two has been a problem.

In 1920 Raymond Franzen proposed a solution that would require little of our attention were it not that, in spite of repeated demonstrations of the unsound assumptions upon which the method is based,⁸ it has proved to be one of the most persistent die-hards in the history of educational psychology. The theory suggested by Franzen was that inasmuch as both mental-age standards and educational-age standards are calibrated in such a way that the child who is exactly average in both measurements will have an educational age as well as a mental age corresponding to his chronological age, with both IQ and EQ falling exactly at 100, the ratio between mental and educational ages (or, what is the same thing, between EQ and IQ) may be taken as a measure of the extent to which a child is "working up to his ability." To this ratio, (EA/MA) or (EQ/IQ) , Franzen gave the name of the Accomplishment Quotient (AQ) or, as it is often called, the Accomplishment Ratio (AR).

The idea appealed at once to teachers and school administrators. Now at last it appeared that a method had been found by which lack of effort could be distinguished from lack of ability, not only in a general way but by actual quantitative measurement. Bright children could no longer be looked upon as doing satisfactory work merely because they were up to the average of their classes. Poor teachers could no longer

⁸ Franzen himself later called attention to some of these errors.

cite the low mental status of their pupils as an excuse for the inferior quality of their achievement. The accomplishment ratio of the class would provide an objective measure of their teaching proficiency in terms of what their pupils might fairly be expected to accomplish. Small wonder that the method immediately became popular or that school people have been loath to relinquish it!

There are, however, at least three sources of error in the procedure, and unfortunately these errors are likely to reinforce rather than to cancel each other both for individual cases and in group measurements. These errors may be designated as follows: those arising from unequal placement of the zero point in the two measures, those resulting from unequal variability, and those caused by failure to allow for regression due to errors of measurement.

The first of these may readily be understood by the following example. Suppose we are dealing with three children, each of whom has a mental age of exactly nine years. If the theory underlying the accomplishment ratio were correct, all should be doing schoolwork of equal quality. Their educational ages should be exactly nine years. But suppose that their intelligence quotients are respectively 150, 100, and 75, which means that their chronological ages would be six, nine, and twelve years. In order to justify the assumptions of the AR, the first child would have to acquire overnight⁹ that which the second child has taken three years and the third six years to learn. That bright children do learn more rapidly than backward ones is unquestionably true, but it is putting a considerable strain on them to ask that they be able to perform such feats as this.

The factor of unequal variability is just a special case of that which was discussed in connection with the intelligence quotient. It was pointed out in Chapter 11 that unless the standard deviation of IQ's is equal at all ages, the IQ cannot have equal significance at all ages. An IQ of 140 at one age will indicate no greater degree of superiority than one of 120 at another age if the S.D. of the first distribution is twice as large as that of the second. Now if, as is usually the case, the variability of the educational ages and the mental ages used in computing the accomplishment ratio is not the same,¹⁰ then a correction for this factor must be made if the comparison between the two is to be valid. Such a correction is entirely feasible if the facts are known, but the ordinary

⁹ Assuming that all entered school at the age of six and that none received instruction in the school subjects prior to school entrance.

¹⁰ Although no general rule can be laid down since tests differ so greatly in this respect, it is usual to find that the S.D. of the educational ages earned on a well-standardized test is smaller than that of the mental ages obtained by the same subjects on a good intelligence test.

method of computing the accomplishment ratio makes no provision for it.

The third point is merely a simple application of the principle of regression. The most likely prediction of the score on one fallible measure that can be made on the basis of that earned on another fallible measure depends on (a) the degree of correlation between the two measures and (b) their relative variability (see page 206).

For all these reasons, the accomplishment ratio in its usual form is not a device to be recommended. If a comparison between the results of educational tests and intelligence tests is desired, the regression formula given on page 206 can be used to find the most probable standing on the educational test corresponding to any given score on the intelligence test earned by children of a given chronological age. The standard error of this estimate will be equal to $S.D._{int.} \sqrt{1 - r^2}$ where $S.D._{int.}$ is the standard deviation of the mental ages or IQ's used to predict the educational standing and r is the correlation between the two measures. If the difference between a child's obtained score on the educational test and his predicted score on that test is significantly greater than chance, as indicated by the error of estimate, would lead one to expect, then the assumption that he is doing either less well or better than the average child of his age and level of intelligence may be warranted. In either case, many explanations are possible. If the educational level is significantly below expectation, the child may be suffering from some unrecognized physical defect or emotional tension, or the trouble may be simply lack of interest and effort. If the educational achievement exceeds expectation by a significant amount, he may have developed unusually efficient study habits or may be driven by over-ambitious parents to spend more than the usual amount of time at his lessons.¹¹

When the comparison of educational achievement with intellectual level is to be made for a group of children rather than for a single child, the usual procedure for studying the dependability of a difference between the scores on two correlated measures may be used. (See Chapters

¹¹ Because of the fact that the AR is frequently called a "measure of the extent to which a child is working up to his ability" those who accept this definition are frequently puzzled by the fact that many children earn AR's above 100. For if an AR of 100 means that a child is already working up to the level of his ability, how can such a performance be exceeded? In a recent textbook on mental and educational measurements, several paragraphs are devoted to this question. The author finally concludes that such a thing is manifestly impossible, and that the apparent instances of its occurrence must be due to inaccuracies in the tests used or in the manner of using them! Of course the whole question is absurd. An AR of 100, even if it were a legitimate measurement, would not signify maximum but only average achievement for a given mental and chronological age.

15-16.) Or, if one is willing to accept the rather dubious hypothesis that the mental and educational ages of a child should not differ by more than the amount that can reasonably be accounted for by the standard errors of measurement if he is doing as well in school as his intellectual level would lead one to expect, then a direct comparison of his standard scores on the two measures with appropriate allowance for errors of measurement will suffice.

PRACTICAL APPLICATIONS OF EDUCATIONAL MEASUREMENT

The comparative specificity of educational tests gives them many advantages over most other measures of mental functions. The boundaries of the universe which we call *achievement in reading* are by no means sharply established, it is true. Reading is not a unitary skill but depends upon a number of underlying factors such as rate and span of visual apprehension, size of vocabulary, level of general information, and the like. Nevertheless, few will question that there is much closer agreement from observer to observer as to what constitutes reading ability or arithmetic ability than is likely to be found with respect to such characteristics as introversion-extroversion, intelligence, or emotional stability. Although educational measurements have not attained complete objectivity of definition, they have approached much more closely to that ideal than is true of most other types of mental measurement. This fact gives to the test user a much clearer idea of just what he is measuring than would otherwise be possible.

The specificity of educational tests is mirrored in their generally high self-correlations. Most of the well-standardized educational tests of the performance type (see page 325) have shown correlations between retests on the same form or between equivalent forms after a short time interval that run as high as $+ .90$ or above for representative samples of subjects of the same chronological age. The self-correlation of product scales and of information tests on the school subjects is usually not quite so high unless the average rating of several judges is used for scoring those of the former type and the size of the sample of information items is large. Even in these cases, however, the self-correlations of the better-standardized scales are higher than are usually found for most other types of mental test.

This self-correlation imbues the test user with a feeling of confidence in the results of his measurement. He knows what these measurements mean, at least within reasonable limits, and is thereby prepared to put his findings to practical use in the immediate handling of children or

in the solution of scientific and educational problems. He may use them for the classification of children into ability groups, or as an aid to deciding doubtful questions of promotion. He may use them as criteria when studying the effectiveness of different methods of instruction or other questions of educational procedure. By comparing the relative performances of selected groups with respect either to mean performance or to variability of performance on the various tests, considerable light may be thrown on such questions as racial or sex differences, the effect of certain physical and sensory handicaps, bilingualism, and so on. Finally, a comparison of the relative proficiency of a given child in the different subject fields, granting that he has had adequate opportunity to acquire such proficiency, is an important first step in the study of specialized individual talent and of the factors which underlie it. This topic will be considered in the next chapter.

The Measurement of Special Talents and Deficiencies

THE RELATION OF SPECIAL TALENT TO GENERAL INTELLECTUAL ABILITY

It will be recalled that Spearman defined special abilities as the residual factors underlying a correlation matrix after the general factor (g) has been rendered constant. The layman sometimes thinks of special talent in terms that are not wholly unlike Spearman's concept of it. He does not use the same standard for all people but makes allowance, within rough limits, for the subject's general level of competence in considering the fields in which he is most talented or deficient. The layman, however, is unlikely to be entirely consistent here. His judgments waver as his standards shift from the individual to the group. At one moment he is inclined to regard John Doe as gifted along mechanical lines because he can make simple repairs of household and farm equipment although he has never been able to learn to read or write or to manage his own affairs without help. But he then recalls that after all, Doe's mechanical ability is hardly equal to that of the average man or boy accustomed to the occasional handling of tools, and that he has never been able to support himself by its use. This realization forces him to qualify his original judgment of Doe's mechanical gifts by some such phrase as "compared to what he can do along other lines" or "does it well for *him*."

The two standards of reference are necessary and important, but they must be kept separate from each other if confusion is to be avoided. The group standard in which the individual is compared with others of his class must be used whenever social or industrial competition is involved. The individual standard is used in determining the lines along which a given subject is most capable or most lacking in competence. In practical work, both the individual standard and the group standard may be used.

The nature of the individual standard is not as clear cut as is the group standard. Actually it is neither a single measure nor a statistical composite of many. It is a global view of all that is known of the person in question, with his general intellectual ability occupying a key position though it is not the only factor considered. As far as the results of tests and measurements are concerned, a rank-order comparison of the result is usually all that is required.

METHODS OF STUDY

When an individual is to be compared with the group to which he belongs, the main consideration is that of making sure that the character of that group has been properly established and that the normative standards are adequate both with respect to means and to variability. Such factors as age, sex, and general experience in the field measured must be carefully considered. If the test used is not a direct measure of intelligence but is one in which differences in intelligence play a part in determining the score, this fact should be taken into account in the interpretation of individual scores. If intelligence is an asset for the skill in question, equal test scores made by two persons may be due, in the one case, largely to superior intelligence; in the other case, special aptitude for the particular skill overtly measured may play the leading role. This does not necessarily invalidate the measure used, for the equal test scores may still predict equal performances, even though the factors determining the performances differ. But it is likely that, had circumstances led that way, the subject with the higher general ability might have done equally well along some other line, while the second, who owes his high test score chiefly to specialized talent for the field in question, is probably more circumscribed as to the fields in which success is likely.

When an analysis is to be made of the pattern of abilities shown by any particular person, a graphic representation of his standing along each of the lines measured is probably the most effective way of showing the facts. By transmuting the results of each test or other type of measurement into comparable units, it not only becomes possible to show the individual's relative standing on each of the various measurements, but also to indicate how he compares with others of his group. A diagram of this kind in which both intrinsic and extrinsic standards are applied to the diagnosis of individual abilities is called a *psychological profile* or a *psychogram*.

The first step in the preparation of a psychogram consists in reducing to a uniform system of measurement all the facts that are to be shown

on it. From what has been said in previous chapters it should be apparent that many of the methods that have been used for this purpose are not properly applicable. Neither age standards nor percentiles represent equal linear distances along a quantitative scale; yet each has not infrequently been used for constructing psychograms where equal calibration is required, with resultant misleading effects. Quotients based upon age standards partake of their inequalities of calibration and involve further hazards of their own. The most suitable unit for the purpose is undoubtedly the standard score, and since the experimental errors of the measures to be compared are likely to differ widely in magnitude, it is wise to substitute the estimated true standard scores on each of the measures in place of the obtained scores. Reference to the formula for estimating true scores given on page 166 will show that when standard scores are used in place of raw scores, the final term becomes zero. The best estimate of the true standard score is then

$$r_{11}(x),$$

where x is the standard score obtained by dividing the difference between the subject's score and the mean of his group by the standard deviation of the group, and r_{11} is the self-correlation of the measure.

Figure 24 illustrates the use of the psychogram for presenting the results of a series of intelligence tests and tests of educational achievement given to a twelve-year-old junior high school boy.

The heavy vertical line in the center of the figure indicates the average score for children of his age on each of the tests given. The lighter vertical lines on either side mark off points located one, two, and three standard deviations from the average. Those to the left of the central line indicate points below the average; those to the right of it indicate corresponding levels above average.

A glance at the figure shows the nature of this boy's difficulty at once. He is an exceedingly poor reader. This deficiency is indicated not only by the fact that his scores on the two reading tests are markedly below average in spite of his good general intelligence, but by the pattern of his performances on the other tests as well. On the three intelligence tests his best performance is that on the Cornell-Coxe, a nonlanguage test not requiring reading. His poorest is on the Terman Group Test, in which the poor reader is at a considerable disadvantage. On the two reading tests, he does relatively better on the one involving the smaller units (word meaning) than on the one which demands reading in larger segments (paragraph meaning). Of all the educational tests given, he does best in arithmetic computation. The test of arithmetic reasoning again requires reading, and as it is a timed test the

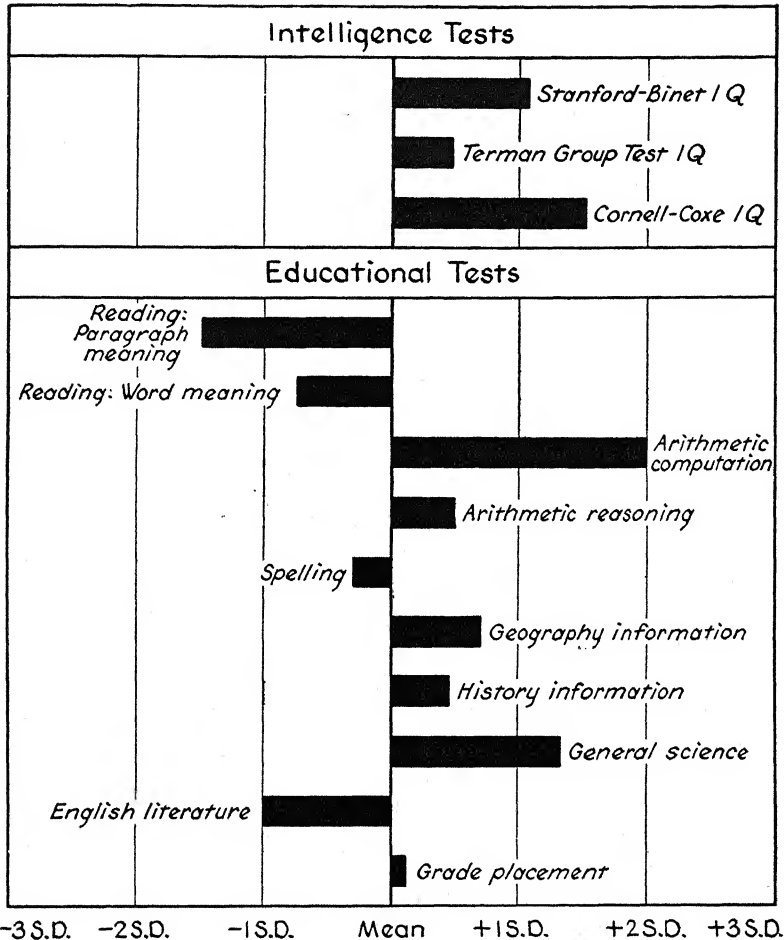


FIG. 24. PSYCHOGRAM SHOWING THE STANDING OF A POOR READER ON TESTS OF INTELLIGENCE AND OF SCHOOL ACHIEVEMENT.

chances are he was obliged to spend too much time in reading the questions to show what he was really capable of doing had more time been allowed.¹ The scores on the tests of geographical and historical information are rather surprisingly high in view of the fact that both tests

¹ Further evidence in support of this conclusion is found in the original papers. Although the number of problems correctly solved on the reasoning test was fewer than would have been expected from his high score in computation, no errors were made on those attempted. This suggests that his slow rate of reading the problems was largely responsible for the lowered score, since he apparently solved correctly all those he was able to read within the allotted period of time.

demand reading, but the school records show that the boy's interest in both these subjects is very keen. The general science test includes a number of pictorial problems for the solution of which his reading difficulty would not be much of a handicap.

In addition to tests of ability, psychograms sometimes include records of physical measurements; personal-social characteristics; habits, such as average number of hours of sleep or length of home study a day; environmental factors, such as measures of the home and neighborhood; family income; educational level of parents; and other facts for which some form of quantitative record can be obtained. When these are included it is desirable to group the material according to type of measurement, putting measures of intelligence in one section of the diagram, educational measures in another, and so on for each general area included.

Because of the relatively precise nature of the data included, the psychogram is particularly useful in the study of the educational achievements of individual school children. When the method is extended to clinical work with children showing behavior problems of various kinds, and an attempt is made to employ the so-called tests of personality and character in the same way, the questionable validity of many of these tests must be taken into account. The use of estimated true scores in place of original scores makes allowance for lack of stability in the test results but it does not correct for the fact that the test may not measure that which it purports to measure. There is more danger that this fact may be overlooked when the data from tests of questionable validity are entered on the same graph with those for which the general appropriateness of the name by which they are designated has been reasonably well established.

As a part of the clinical case study, the psychogram has a definite field of usefulness if the test results and the observational material are brought together in such a way that each serves to interpret and amplify the other. For example, in cases where the test scores in a particular area deviate significantly from what might be expected on the basis of those earned on the remainder of the test series, the case history material may sometimes provide cues which will either lead to an explanation of this fact or modify the interpretation that might otherwise be made from it. Low reading scores are understandable if there is a history of serious visual defect which was not corrected until after the child had passed into a grade where the reading had become so difficult that he could make no progress with it. Poor schoolwork may be the only weapon at the command of a child who bitterly resents the inflexible demands of parents who are overambitious for his success. Emotional distress over family

discord and the threat of a broken home may seriously disrupt the school progress of a sensitive child.

Just as the case history record may throw light on unexpected aspects of the child's test performance, so the tests may help to explain the case history findings. Contrast between the intellectual level of the child and that of the other members of his family can account for much home friction. That this is likely to be true when the child's IQ is definitely below that of his parents and siblings is easy to understand, but the bright child in the family of mediocre or inferior intellectual gifts is also liable to meet difficulty. His preoccupation with books and reading which diverts him from more "practical" affairs, his attempts at scientific experiments which sometimes come to grief, and a host of other differences, many of them trivial in themselves but significant in the aggregate, exasperate the parents and disturb the child.

Educational deficiencies do not cease from troubling when the child passes the door of the schoolroom. They follow him to the playground, where his companions boast of high marks, and his only recourse appears to lie either in bravado or in lying. They follow him to his home, where the repercussion may take the form of scolding which he resents, attempts at help which keep him from play, or sympathy which undermines his courage and self-respect. Behavior difficulties of some form, the source of which often seems as inexplicable to the child as to the parents and teacher, are likely to follow. They all know that he is not doing well in school. They know that he seems morose and sullen, loud-mouthed and boastful, rebellious, irritable; or it may be that he appears nervous, worried, and depressed. But they do not connect the behavior with the educational handicap; or, if they do, the chances are they will consider the behavior as the primary factor, the poor schoolwork as its result. Actually, in most such cases, the reaction is circular, each having an adverse effect upon the other. But the specificity of the educational handicap makes it the easiest point to attack, and in many cases an improvement in school accomplishment will bring rather surprising results in areas that at first thought seem to have but slight connection with it. Arthur (1946) has given a very clear account of the favorable effect upon general behavior and adjustment that may often be brought about by successful tutoring of children whose school difficulties were erroneously believed to be the *result* of misbehavior and lack of interest when the causal relationship was mainly in the opposite direction.

An intelligent approach to child guidance demands knowledge of the facts. The more clearly these facts are organized and marshaled for inspection, the more likely it is that fundamental relationships will be seen. The psychogram is one way of doing this. By using estimated true

standard scores in place of uncorrected standard-score values, appropriate allowance is made for differing self-correlations ("reliability") of the measures used. This, however, will not correct errors of interpretation arising from the "naming fallacy" when tests of dubious significance are included in the picture. If such tests are used, it is well to set them off from the others in such a way that the confidence which the better tests merit will not carry over to those of questionable meaning merely because of proximity.

MEASURES OF SPECIAL TALENT

The widely differing concepts of talent held by people in general is indicated by the fact that in *Webster's new international dictionary* more than half a column is required to define its various nuances of meaning. In Warren's *Dictionary of psychology*, however, only the following appears: "Talent = a natural aptitude which renders the possessor susceptible to a high degree of training in some special field of endeavor, such as music, diplomacy, etc." Warren's definition does not appear to exclude the tool subjects of the elementary school except, perhaps, by implication in the use of the term "field of endeavor," but most people prefer to limit the use of the word "talent" to areas in which at least the gifted few find their lifework and which demand a high degree of special skill. Among these, the fine arts occupy an important place.

One of the earliest attempts to develop an objective way of identifying musical talent in advance of special training was made in 1919 by Seashore of the State University of Iowa. Seashore believed that musical ability is susceptible to analysis into a number of simpler elements of which the following are among the most simple and essential: sense of pitch, sense of intensity, sense of rhythm, sense of time, sense of consonance, and tonal memory. A series of phonograph records was prepared to which the subject listened. Each record contained a series of paired notes or musical phrases. The subjects were required to state in each case whether the second note was higher or lower, louder or softer, or whether two phrases were alike or different, and so on, according to the particular musical factor to be measured.

Seashore's tests attracted much attention and have been widely used in schools, colleges, and conservatories of music. A good many investigations have been conducted to determine the stability of the results as indicated by repeated measures of the same subjects, as well as to ascertain the extent to which the tests will distinguish between students who do well in their music courses and those who do poorly. The findings have not been entirely uniform but in general they have indicated

only moderate agreement between the results of tests and retests, especially when the two are separated by some length of time. The relationship between test scores and ability to profit by instruction in music is not rectilinear. Those subjects who do very poorly on the tests are rarely able to become more than very mediocre performers, even under long tutelage and faithful practice. But exceptionally high scores appear to be more closely related to keen sensory perception than to high musical talent. It would appear that a reasonably good "ear for music" is essential for superior musical achievement but that auditory sensitivity alone does not guarantee success along musical lines. Something more is needed.

Since the appearance of the Seashore tests, many other people have attempted to develop tests of musical talent or to devise ways of measuring the level of accomplishment along musical lines. The former have in general followed Seashore's plan of using phonograph records. In a good many instances, these persons have also adopted Seashore's rather molecular view of musical talent, and have based their measurements upon part-processes, rather than attempting a more global view. Others have turned to tests of musical appreciation, feeling, and understanding.

The measurement of progress in musical education is in many ways an easier task than that of measuring talent directly and in advance of training. Inasmuch as when training and practice are kept equal, the rate of musical progress is to some extent an indication of talent, it may well be that until more useful methods of measuring talent in advance of training are developed, progress records may prove to be the most useful prognostic devices for practical use that are available at the present time. The major difficulty here lies in the fact that, especially when children are concerned, interest plays so large a part in determining progress. Some teachers are much more apt at arousing and maintaining the interest of their pupils than others are. For this reason, lack of satisfactory progress under one teacher at a particular time is not valid evidence that more rapid advancement would not be made under another teacher later on. Here, as in many other instances, positive evidence is more conclusive than negative evidence.

Tests of musical achievement include such matters as sight reading, information tests, ability to recognize musical themes and musical compositions, the reproduction and memorizing of musical phrases and rhythms. Tests of the human voice are sometimes made by the use of formal instruments such as the tonometer or the tonoscope, but because there are many subtle differences in voice quality which these instruments do not record, judgments of qualified listeners are generally preferred when something more than mechanical determination of the accuracy of pitch and similar characteristics is in question. Some attempts have

been made to produce greater objectivity in these judgments by the use of formal rating scales.

Interest in developing tests of musical talent, appreciation, and achievement has been very marked. In the two editions of Hildreth's *Bibliography of mental tests and rating scales* (1939, 1945) 86 titles are listed. This does not include the many studies of the reliability and validity of these tests.

Attempts to identify artistic talent in advance of its manifestation, like the corresponding attempts in the field of musical talent, have been more successful at the negative than at the positive end of the distribution. It is easier to identify those with little ability along artistic lines than those who have much.² For true artistic talent enables its possessor to do more than produce a photographic reproduction of the scenes which he depicts. He selects, combines, strengthens this feature and softens that, and by so doing creates a feeling tone and clarifies its meaning. Löwenfeld (1939) differentiates the "optic" type of graphic or plastic expression, in which the subject attempts to reproduce an object or scene as he sees it, from the "haptic" type, in which the reproduction is in terms of what it means to the artist, the emotional response which it arouses in him and which he, in turn, tries to convey to others. In the belief that in the artistic productions of the blind and weak-sighted would be found one of the clearest demonstrations of this principle, Löwenfeld made an extensive collection of the drawings, paintings, and sculpture of persons with very defective vision, including a few who were totally blind. These were compared with the spontaneous drawings of young normal children and those of primitive races, as well as with those of older subjects with normal vision. His analysis of the similarities and differences in the work of these groups is vividly illustrated by several hundred excellent reproductions. The figure of "Youth Imploring" (see Figure 25) affords a beautiful example of the difficulty of devising a wholly objective test of artistic merit. If judged by the "optic" standard, that is, by the degree of conformity of the product to objective reality, the statue would seem to possess little merit. The proportions are wrongly conceived and the whole is none too well executed. But these very defects are what give the figure its extraordinary quality of feeling. The exaggerated length of the arms and size of the hands, the upward strain of the chest and shoulders contrasted with the humble kneeling posture combine to give it a high rank according to "haptic" standards.

Although a fairly large number of attempts have been made to

² Assuming, of course, equality of training and opportunity.

measure artistic talent, and even more effort has been expended upon the measurement of artistic appreciation and on achievement tests designed to measure progress in artistic production, none of these can be said to have met with a high degree of success. If artistic talent is based on a combination of more elementary abilities, no one has succeeded in



FIG. 25. "Youth Imploring," A CLAY STATUE MODELED BY A CONGENITALLY BLIND YOUTH OF SEVENTEEN YEARS. (Reproduced by permission of the publishers from Plate 35, p. 232 in *The Nature of Creative Activity* by Viktor Löwenfeld. New York: Harcourt Brace and Company, 1939.)

isolating these underlying factors experimentally, though several have attempted to do so by means of armchair logic. At the State University of Iowa, Meier (1941), as the result of many years' study of the problem, has come to the opinion that artistic ability is "a complex of abilities and functions that are peculiarly interlinked." Of the six factors which he names, three, so he thinks, are primarily due to inheritance. These are (1) manual skill or craftsman ability, (2) volitional per-

severation, and (3) aesthetic intelligence. The remaining three refer more directly to learning. They are (1) perceptual facility, (2) creative imagination, and (3) aesthetic judgment. He quotes several studies to show that artistic talent tends to run in families and, from such biographical data as are available, attempts to show that the three characteristics first named were predominant in the ancestry of famous artists, while the latter three were stimulated by the quality of their individual experience. Unless the constitutional tendency is present, however, experience will be relatively ineffective, while in the absence of suitable stimulation the outlook that latent talent will become manifest is slight. Although such an analysis is plausible enough, it needs confirmation by more objective methods. A factor analysis of the results of a number of the best tests of art talent, art appreciation, art judgment, and drawing achievement might throw light on the problem.

Methods employed for the study of art talent vary greatly. The earlier devices were for the most part product scales (Thorndike, 1913; McCarty, 1924; and others) or rating scales. Later attempts have included tests of art judgment such as that by Meier and Seashore (1929) in which the subject is required to choose the better of two pictures, one of which violates some recognized principle of composition or design while the other is conventionally correct in that respect. Tests of artistic information, aesthetic appreciation, and color harmony have also been used, and some attempts have been made to develop performance tests to measure, according to some such formal analysis as that given above, the several elements believed to underlie art talent. None of the latter studies, however, have been as well standardized as were the early product scales, which were fairly serviceable for judging the merit of simple drawings made by children in the primary grades. However, their relatively formalized nature does not conform to modern concepts of art education. For this reason they have been relatively little used during recent years.

Talent for writing English prose or poetry is like talent for the other fine arts in that its formal aspects are comparatively easy to measure; but the upper levels, where mere correctness gives place to that indefinable tone and flavor which we call "literary quality," are not easy to define or appraise. The available tests include product scales for judging the quality of written compositions, tests designed to measure knowledge of the rules of literary construction, and tests of literary appreciation, in which the subject is required to select the better of two selections of prose or poetry or to arrange a series of such selections in order of merit. The question of a criterion, in these cases, has caused some difficulties. Abbott and Trabue (1921) solved the problem by

taking as their standard some of the less well known selections from Shakespeare, Tennyson, and other poets and pairing these with compositions of their own devising. Some of the latter were of the "sickly sentimental" variety; others were trite and formal. Although all were written in correct poetic form and the general subject matter was the same as that of the original, the poetic flavor of the latter was intentionally destroyed. This test is undoubtedly one of the best of its kind and merits wider use than has been accorded to it.

Talent for social relationships undoubtedly exists and some attempts at measuring it have been made. These will be considered in Chapter 26. Talent in areas with a more directly vocational aspect such as clerical ability, business administration, or salesmanship, as well as special aptitude for the mechanical arts, will be discussed in Chapter 27.

CRAFTMANSHIP VERSUS CREATIVENESS

To *create*, according to *Webster's new international dictionary*, is "to produce . . . a work of thought or imagination . . . along new or unconventional lines."³ The great men of history were more than good craftsmen who patiently reproduced the patterns set by their predecessors. They brought something new into the world—a new scientific discovery, a new social plan, a great work of art, music, or literature.

If we could identify while they are yet children those capable of becoming the geniuses of the future, and provide them with the kind of training and opportunity for self-development best suited to their needs, the advantage to future generations would undoubtedly be very great. Terman, in his studies of intellectually gifted children, has laid the groundwork for this task, but much more is wanting for its complete fruition. The basis of selection in Terman's studies was an IQ of 140 or over on the 1916 Stanford-Binet. More than a thousand children ranking at this level were chosen for study, and their progress has now been followed for well over a quarter of a century (Terman and Oden, 1947). With few exceptions, the level of accomplishment attained by these subjects during the early years of their maturity has been much above the average of the population as a whole, but to date there has been none whose achievements would entitle him to first rank. They are successful in business and in the learned professions; they occupy positions of trust in the communities in which they reside; their incomes are for the most part good and in a few cases very high. The method seems to have proved successful in selecting children who, as adults, occupy the top

³ This, of course, is not the only meaning of the word, but it is the one that best fits the present discussion.

but not the topmost rank among young American citizens. Perhaps as time advances, some will attain this coveted place, for all are still young when compared to the age at which men commonly achieve fame. At the time of the 1947 report, practically none had reached the age of fifty years; many were still in their thirties. There is still time for some among them to achieve the rank of true genius.

What is genius? Is it only an exceptionally high level of "general intelligence" of the kind measured by the best of our present-day intelligence tests? Is it specialized talent of a correspondingly high order? Is it merely "an infinite capacity for taking pains?" Is genius of a particular order coextensive with craftsmanship along the same order? Or is it something derived from and yet more than any of these, an interaction among them which gives rise to a unique pattern of thought and a type of response that derive their power through the reinforcement of each of the elements by all the rest?

No final answer to these questions is possible from current knowledge and experimentation. Nevertheless, the failure of our present methods to identify and measure abilities beyond the level of superior craftsmen strongly suggests the existence of some additional factor that lies beyond our grasp. It may be, as was suggested in the last paragraph, the pattern of mental organization, the type of interaction between abilities, the motives which lead to their overt manifestation, and the conditions under which these manifestations occur. Or it may be that creative imagination is as much a distinct ability as is artistic or musical talent.⁴ If the former should prove to be the case, some form of statistical analysis would seem to be the most promising avenue of approach. If, on the contrary, true creativeness depends upon a special kind of intellectual gift, if it belongs in the class of specialized aptitudes or talents, then a new and different type of measuring instrument would seem to be called for. Although a few attempts to develop tests of this kind have been made, they have not proved sufficiently useful to warrant special consideration here.

"IDIOTS SAVANTS"

From time to time, considerable interest has centered about the occasional cases of unusual ability in some special field that appear among those whose general intelligence is of a very low order. Almost any institution for the feeble-minded can show one or more instances of

⁴ There is evidence that all mental traits are correlated to some extent, at least as far as the extreme cases are concerned. In this sense, no one of them can be looked upon as wholly distinct from the others.

patients who, if not highly gifted as judged by the standards applied to normal people, yet exhibit some talent which appears remarkable in view of their deficiency along other lines. In the institution at Faribault, Minnesota, for example, there is a patient of low-grade imbecile level who can reproduce any simple melody on the piano after a single hearing. She also knows a good many songs and marches which she plays for her own amusement and that of the other patients in her ward. She does not know the words of the songs, and her speech in general is quite rudimentary, as are her other accomplishments.

Early in the present century, a Negro imbecile known as "Blind Tom" aroused much attention on the vaudeville stage. Blind Tom's accomplishment was similar to that of the child just mentioned except that it functioned at a higher level. His repertoire included a good many rather difficult selections, and his performance of them was more finished. Like her, he could reproduce selections heard but once, and it is said that if an error was made in the original performance, that, as well as the portions correctly played, would be repeated with machine-like fidelity. According to report, Blind Tom was never able to learn to dress himself, and his eating habits and other behavior were uncouth in the extreme.

Two cases that have come under my own observation may be mentioned here. The first was a little Italian boy of about twelve years of age whose Stanford-Binet mental age was approximately four years. This boy's talent for mimicry was so marked that had he lived to adult years, he might well have become a success on the vaudeville stage. His imitation of a cat fight was so true to life that stray cats who heard his caterwauling would come rushing to what they judged to be the scene of action. Other animals as well as people were imitated with equal success. A lame man with a singularly dour countenance and shrill, quavering voice who lived in the neighborhood was aped so perfectly that no one who had ever seen the original would fail to recognize the copy. In this case, the talent went far beyond that of most people. It was not merely a matter of contrast with the child's general level of ability as is true of the case next to be described.

Bertrand was a boy of eleven years who had never attended school until a special class for low-grade feeble-minded children was established as an experimental project by the public schools of the city in which he resided. Bertrand's mental age was less than three years. His gait was shambling with partially bent knees, swinging arms, and stooping posture. His speech was slurred and indistinct, and he drooled constantly. Nocturnal bladder control had never been established and there were occasional accidents during the day.

In school he seemed at first to make no progress along any line. Even the simplest manual occupations were beyond his mastery. He was a handicap to the play of the other children, even though their games were of the simple kind necessitated by an average mental age of around four years. The period in the school day which he chiefly enjoyed was the story hour. Then he sat motionless, utterly enthralled by the adventures of "The Little Red Hen" or "The Old Woman and the Pig."

A few of the more advanced children in the class had a brief daily lesson in reading. One day, when the teacher was conducting a drill in word recognition by means of cards on which single words were printed in letters large enough to be seen across the room, she was amazed to note that Bertrand had joined the group and was pronouncing the words along with the others. An individual test disclosed that he actually knew a number of the words. From then on he continued to join the other children for their reading lesson, and by the end of a year had progressed to the point where he was able to read an easy primer without much difficulty. By the end of two years he could read anything that his limited understanding enabled him to grasp—which indeed was not much. But he reveled in such nursery tales as "Little Black Sambo" and "Peter Rabbit" and his delighted parents provided him with a large supply of these infantile classics.

Bertrand's phenomenal progress in reading—for such a performance *is* phenomenal in a child of his mental level—was not paralleled by similar advance along other lines. He never learned to dress himself without help, nor did he ever gain complete control of his bladder. His table manners were unpleasant, for his clumsy hands frequently dropped food or upset liquids. Choking and dribbling occurred frequently. His concept of numbers never advanced beyond that of "one," which was usually named correctly, and "more than one" which was generally designated as "two," although he could "count" in the sense of repeating the numbers as far as ten and could read numbers up to one hundred.

A fair number of such cases of exceptional talent along a single line among children or adults of defective intelligence have been reported in the literature. The "genius of Earlswood Asylum" described by Tredgold (1908) is one of the earliest; that by Scheerer, Rothmann, and Goldstein (1945), one of the most recent.

Explanatory theories for these cases fall under a number of general heads. Some are inclined to regard them as instances of organic brain damage in which certain areas—those mainly concerned with the type of ability manifested—either escaped injury or suffered less pronounced damage than the remainder of the organ. Allied to this is the belief that

many if not most of such cases are in reality psychotic, rather than mentally defective, and that their peculiar patterns of ability and inability are merely special manifestations of the psychotic personality. A third theory, suggested particularly by Binet in his *Psychology des grands calculateurs* (1894), is that the presence of some initial special aptitude together with relative deprivation of success along other lines, and with environmental circumstances fostering practice in the field of the special ability may combine to bring about extraordinary facility of performance in the single limited area. Binet describes the case of a young man of very limited general intelligence who had spent his boyhood and youth in the darkness of a coal mine, where his only task was to open and close a door through which the cars loaded with coal had to pass. In his unoccupied periods he fell to practicing numerical calculations. In time he developed very remarkable facility. His success was due in part to the discovery of certain short cuts and mnemonic devices with which people in general are not familiar, in part to natural aptitude, and in part to special interest and effort leading to an exceptional amount of practice. This was fostered by the conditions of his work, which left him with much idle time in which other occupations were impossible, and by the wonder and acclaim with which his performances were received by others, which provided a powerful source of motivation for him.

Scheerer, Rothmann, and Goldstein propose a slight modification of Binet's theory. By means of an extensive series of tests and measurements they were able to show that the boy whom they studied showed a general and widespread impairment in practically every type of abstract thinking. This impairment extended to his use of language symbols, to his concepts of number, social phenomena, and other aspects of functional relationships, and reduced his Binet test performance to a level represented by an IQ of 50. But he could name the day of the week corresponding to any given date between approximately 1880 and 1950. He could also perform other highly specialized tasks but with little regard for their conceptual framework. From a study of this case and that of others reported in the literature, the authors come to the conclusion that

a defective organism will cling tenaciously to those aspects of a situation and those features of material which make concrete palpable sense to him, i.e., with which he can deal successfully. These aspects are thrust into the foreground of the phenomenal organization as the "figure." Such a difference in perceptive centering in the abnormal's coming to terms with the world of the "normal" leads to a different centering of performance. Therefore these aments retain easily what may appear senseless or peripheral or irrelevant to the normal

observer. To the aments in question, however, this is the only "sense" possible and pivotal in the experienced contents.⁵

SPECIALIZED MENTAL OR EDUCATIONAL DEFICIENCIES

In considering cases of special mental or educational defect it is necessary to make a sharp distinction between those forms of inability that are associated with, and may be considered the direct or indirect results of general mental backwardness, and those in which the lack of ability is limited to one or more single areas which are much out of line with the subject's general level of performance. Just how great the difference must be to constitute a "special defect" is a matter of opinion about which no general agreement has been reached. Perhaps more rigid standards should be applied to some areas than to others. Certain kinds of defect seriously handicap the person who possesses them. These handicaps extend far beyond the immediate field in which the inability is displayed. Others, although equally pronounced, are unlikely to cause really serious inconvenience. Some people are unable to distinguish the pitch of tones; they are tone-deaf. Although they are barred from experiencing the aesthetic pleasures that others find in listening to music,⁶ this loss is unlikely to occasion any real disturbance in their lives or work. But the person who has a serious speech defect or who reads so poorly that he cannot acquire the formal education needed for the occupational field of his choice is in a different position.

Theories as to the causes of reading difficulties are legion, and the therapeutic measures to which these theories have given birth are equally numerous. One weakness common to most of these theories is that they so frequently try to trace all cases of reading difficulty to a single cause, such as a change in hand preference,⁷ or to some special aspect of the child's habitual mode of perception, such as "motor-mindedness," which makes it difficult or impossible for him to learn by the auditory-visual methods of teaching commonly employed; to some unusual peculiarity of vision; to bad habits of eye movement; or to some other highly

⁵ From "A case of 'Idiots Savants'; an experimental study of personality organization," by Martin Scheerer, Eva Rothmann, and Kurt Goldstein. *Psychological Monographs*, 1945, 58. Whole Number 269, p. 63. Quoted by permission of the authors and the publisher.

⁶ Tone-deaf persons can still enjoy the rhythm of music and find pleasure in dancing or related activities.

⁷ The theory that difficulties in speech, especially stuttering, as well as reading difficulties arise from an enforced change in lateral dominance is due chiefly to Samuel Orton of the State University of Iowa. (See Chapter 24.)

specialized factor. The fact that a child cannot read or that he reads poorly is fairly easy to detect, and there are a number of good reading tests on the market which will indicate with a high degree of accuracy the extent of his deficiency in terms of age or grade standards. These tests are useful in identifying children whose reading ability is below standard; they do not tell the cause of his backwardness.

A good many attempts have been made to develop diagnostic tests by means of which the exact nature of special disabilities in the tool subjects, especially reading and arithmetic, may be analyzed. These include mechanical devices such as motion-picture cameras for photographing eye-movements in reading, or tachistoscopes for determining how much can be seen and recognized in a single very brief period of time. There are many elaborate devices for studying particular aspects of visual and auditory acuity and habits of visual perception, such as the tendency to reverse words or letters which leads to confusion between such letters as *b* and *d*, *p* and *q*, or *p* and *d*, or such words as *saw* and *was*. Reversals also occur within a word, leading to confusion between such words as *loin* and *lion*. Other bad habits take the form of omitting words or parts of words such as prefixes and suffixes, repeating or omitting entire lines as a result of failing to make the correct ocular adjustment at the end of the line. Overmeticulousness in attempting to "see" each word and each letter as a separate entity is a habit that sometimes results from unwise emphasis on the part of parents or teachers who become concerned over what they regard as "carelessness" in reading.

Diagnostic reading tests aim at discovering the particular types of errors to which a child is especially prone, and attempt to locate the bases of his poor reading. They should include, first of all, a thorough-going ocular examination and a test of auditory acuity by means of a standard audiometer and a general health examination by a qualified physician. Data should also be secured with respect to school attendance, with special attention to long absences during the first two or three years when the fundamentals of reading are normally acquired. Certain facts regarding the child's home life and family conditions should be ascertained, such as the use of a foreign language in the home, unusual emotional stresses and family discords, the attitude of the parents toward the child's schoolwork, the amount and type of reading matter in the home. Psychological tests will include both verbal and nonverbal tests of general intelligence and a series of general reading tests involving both oral and silent reading. The silent reading tests should have well-established normative standards and should include tests for the comprehension of large units (paragraphs) and tests of word meaning. An oral vocabulary test should also be given to distinguish between deficient

word knowledge as such and inability to recognize familiar words in print. Both rate and power of silent reading should be determined.

The standardized methods designed to identify particular types of mechanical errors in reading include such tests as the following:

If the two words in a pair are the *same*, write S on the dotted line following them. If they are *different*, write D on the dotted line.

bear—dear moon—noon

pear—pear care—cars

Underline the right word to fill each blank in the following sentences:

John is a good

doy boy poy bay

I have a red apple.

pig dig gib big

One of the words in each of the following lines is not like the others. Cross it out.

came	came	come	came	came
was	was	was	was	saw

In the tests of oral reading the examiner records each error made. These are afterward classified as to type and apparent source of difficulty. The total time required for reading each of the standard selections is also noted and points of long hesitation are marked. While the most accurate means of recording eye-movements is the camera, serviceable results can also be obtained by the so-called "peephole" method, in which the material to be read is mounted on a sheet of heavy cardboard with a small hole in the center through which the examiner observes the child's eye-movements and counts (a) number of fixation pauses per line (or per selection), (b) number of regressive movements. The total time required for reading is also recorded, and special notations are made on a duplicate sheet to indicate where the points of major difficulty occurred.

It was formerly believed that poor eye-movement habits were important factors in the causation of poor reading. More recent studies have indicated that this theory puts the cart before the horse. It is entirely true that poor readers have poor eye-movements during reading, but attempts to improve reading by training directed largely or wholly toward improving the eye-movement habits have not, as a rule, brought encouraging results. On the other hand, improvement in reading skill is almost invariably associated with better eye-movements. There are fewer and shorter fixation pauses and fewer regressive movements. A similar difference can be noted with normal readers when the eye-movements during the reading of easy and difficult material are compared. When reading is easy, eye-movements are comparatively regular

with few regressions and only short fixation pauses. As the difficulty of the reading matter is increased, corresponding changes in the eye-movements occur.

The evidence, then, appears to point to the conclusion that while a study of eye-movements during reading has considerable diagnostic value in determining points at which difficulties occur, this fact should not be interpreted as meaning that faulty eye-movements are a cause of poor reading. Moreover, a study of eye-movements, like the other analytic methods which have been described, can do no more than indicate the kind of mechanical difficulties which the subject shows; they afford few, if any, cues to the causes of these difficulties. In most cases these causes are multiple rather than singular, but there is one factor which is almost universally involved. That factor is lack of practice in reading. Almost without exception, poor readers spend less time in reading for pleasure than do good readers. They read from a sense of duty, with little zest or spontaneous interest in the content. Such interest as they feel in improving their speed of reading is chiefly motivated by a desire to get the disagreeable task over more quickly.

This is another example of the circular relationship between interest and success. Without the motivating force that comes with direct interest in *doing*, skill is acquired slowly and with difficulty. This is why external incentives such as rewards and punishment are generally less effective than the internal drive toward doing a thing for its own sake, though rewards are sometimes useful in getting the mechanism under way in the beginning. Universal rules rarely are found to be true, but with respect to reading skill the nearest approach to a universally sound principle that can be made in the light of our present knowledge is this: *The child who spends much time in reading for pleasure is generally a good reader.*

OTHER DIAGNOSTIC MEASURES

Diagnostic tests in arithmetic are aimed at locating specific points of weakness in which special teaching is called for, such as imperfect mastery of the basic addition combinations or of the multiplication table; poor understanding of fractions or of the meaning and placement of the decimal point; failure to learn arithmetical symbols or the special vocabulary of arithmetic problems. Many children, for example, are puzzled by such expressions as "5 pounds of sugar at 8¢ per pound" or by the unexplained use of such words as "average," "total," "sum," or "quotient." Some children fail to grasp the idea that the wording of a question in arithmetic reasoning has anything whatever to do with the solution of the problem. They merely look at the figures and try to

guess at the way they should be treated. Formal tests of the fundamental processes and of arithmetic vocabulary will generally throw light on the nature of mechanical difficulties, while having the child "think out loud" when attempting to solve problems in reasoning will often reveal the source of his difficulties in that area.

Attempts at locating special difficulties in type of mental processes, such as poor memory, deficiencies in visualizing, flighty attention, or unstable emotional control, as possible contributing factors in educational backwardness are likely to suffer from three weaknesses: (1) too small samples of the ability in question, with consequent likelihood of bias, (2) naïve acceptance of the idea that such factors are *causes* rather than symptoms of more basic underlying conditions, and (3) the naming fallacy discussed in preceding chapters. Many poorly trained clinicians, for example, attempt to make a functional analysis of the results of a Binet test. From an armchair consideration of the surface characteristics of the various items, all sorts of solemn pronouncements are made with respect to the child's pattern of mental abilities. If he fails on the tests of copying designs but passes other tests at the same mental-age level, he is said to be deficient in "visual imagery"; if the most advanced item on which he chances to succeed proves to be a digit-span test, he is said to have an "exceptional memory." No account is taken of the fact that in many cases only the performance on a single task can be cited as evidence for the diagnosis made, nor for the possibility that this performance may have been temporarily influenced by chance circumstances. A little girl who was said to be "notably lacking in the ability to give sustained attention" was later found to have wet her clothing during the course of the test because she was too shy to ask the stranger who gave the test for permission to go to the toilet. Too many conclusions from too little evidence have characterized much of the so-called diagnostic work in the past. It is encouraging that the present trend appears to be toward greater caution. As Josh Billings aptly put it, "It's better not to know so much than to know so danged much that ain't so."

SOME PRACTICAL CONSIDERATIONS

Special talents and defects are not confined to a few exceptional individuals. Everyone displays them in a greater or less degree. Although there is reason to suppose that the unique pattern of mental organization that characterizes the individual in maturity has been built upon the inborn tendencies, the potential aptitudes and ineptitudes handed down by his ancestors, much can be done either to facilitate or to inhibit the manifestation of these tendencies as the child develops. Few of the

skills used in everyday life depend upon a single aspect of ability. Nevertheless, just as a chain is no stronger than its weakest link, so a complex ability may be brought to naught by reason of some specific weakness which, if detected in time, can often be mitigated or corrected. It is here that the value of diagnostic methods is evidenced.

Diagnosis alone, however, is of little value unless something is done about the findings. Concepts of educational and personal therapy have undergone radical changes during recent years. Earlier emphasis was largely upon the external features in the situation, and the methods most in favor involved modification of the environment as changes appeared to be needed as well as external help to the child. More recently there has been increased recognition of the role of the subject in the therapeutic process. Human beings are not inert machines incapable of self-correction or self-help. They are living and growing organisms for whom the keynote of healthy development is activity. And activity is motivated from within. It is not merely a matter of being pushed and pulled about by outside forces. The soundest educational therapy is that which helps the child to help himself.

The Measurement of Motor Development and Motor Skill

THE CONCEPT OF GENERAL MOTOR ABILITY

The success attained by measures of general intelligence and of school achievement has led a number of people to wonder if it may not be possible to devise a scale for measuring general motor ability. Common observation indicates that children of the same age differ markedly in their ability to perform tasks requiring rapid and delicate movements of the hands and fingers, and that they likewise differ in physical strength and in the agility and speed of their bodily movements. Objective evidence, however, has not supported the idea of a generalized motor talent. Unlike the items used in intelligence tests, which regularly show a positive correlation with each other, motor tasks appear to be positively related only when they involve much the same muscle groups, provided always that the subjects used are mentally and physically normal and are of the same age and sex. A person who is awkward and clumsy in his body movements may be exceptionally skillful in the use of his hands. The college athlete who wins all the honors at the track meet may be "all thumbs" when he attempts to thread a needle or to adjust a small screw.

Motor skills of all kinds are greatly affected by practice. Because the various occupations and avocations of daily life differ so greatly in the amount and kind of motor skill required for their performance, it is not easy to find motor tasks of such a nature that reasonably similar opportunity and incentive for previous practice may safely be assumed for all subjects. For this reason, the most successful motor tests have as a rule been designed to measure specific motor skills rather than general motor ability, and have been regarded as measures of achievement rather than of aptitude or talent.¹

¹ Certain motor tests have been found very useful in the selection of candidates for particular kinds of industrial jobs, particularly those requiring rapid and accurate manipulation of objects. It is likely that differences in the anatomical structure of the hands and in neuromuscular coordination as well as differences in experience and practice are involved here.



FIG. 26. EARLY MOTOR SKILLS. (Photographs by Harold M. Lambert.)

THE MEASUREMENT OF MOTOR DEVELOPMENT IN INFANCY AND EARLY CHILDHOOD

Gesell's measures of infant development (1928) include a separate series of motor tests. These involve both gross bodily movements and manual dexterity. No information is given by Gesell as to the relationship between successes on the different items at any single age or on the predictive value of the series as a whole for estimating the most probable

standing of an individual child at some later age on the basis of his performance at an earlier age. Bayley (1935), however, who drew heavily upon Gesell's data as well as upon that of other investigators in the construction of her scale for measuring motor development during the first three years, has provided us with information on this head. Bayley's scale, which consists of seventy-six items calibrated in equally spaced units according to Thurstone's (1925) method of scaling, was given to sixty-one infants at intervals of one month during the first eighteen months and at three-month intervals thereafter.

Bayley found a positive but low correlation between total scores on this scale for the same subjects even when several months had elapsed between testings. As has been found true of mental tests given to infants, however, the correlations between tests separated by only a short period of time were considerably higher. Bayley also obtained a much higher correlation (averaging about $+.50$) between "mental tests" and "motor tests" given to the same subjects at the same time for infants of fifteen months and younger than was obtained for older infants or than has usually been found by other investigators working with older subjects. This, together with other results of her study, led Bayley to question Shirley's (1931) contention that the marked individual differences in motor development shown by the twenty-five babies whom she studied could best be accounted for on the hypothesis of a "motor talent" that different individuals possess in varying degrees. Bayley is of the opinion that *all* developmental items are too closely interwoven during the period of infancy to permit reliable distinction of specialized abilities at these early ages. Children are retarded or accelerated as a whole,² at least as far as measures now available permit us to judge. Later "mental" development can be predicted from early "motor" development as accurately as from tests called by the same name. Even in Shirley's study, tests of prewalking progression did not provide significantly better prognosis of the age of walking alone than could be made from the median age of walking for the group as a whole.

Bayley's study provides the only carefully standardized scale for the measurement of early motor development that is available at the present time. Although her normative standards are based upon only sixty-one infants, the fact that these children were retested at such frequent intervals compensates, to some extent, for the small sampling. There are many studies of growth in specialized forms of motor behavior, of which McGraw's pictorial representations of the successive stages in creeping and in sitting alone (1941) and the valuable study of growth in

² This statement, of course, refers only to behavior, not to growth in physical size.

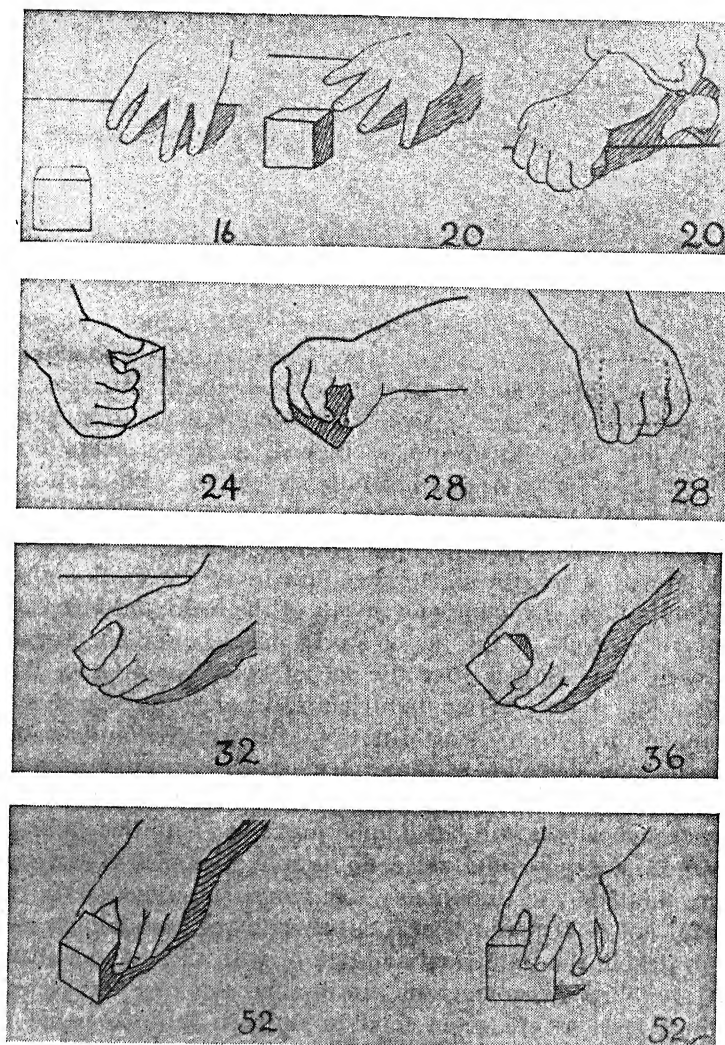


FIG. 27. DEVELOPMENT OF MANUAL PREHENSION IN INFANCY. The ten stages in the development of the ability to seize and hold a one-inch cube have been named by Halverson as follows: (1) reach but no contact; (2) contact but no grasp; (3) the primitive squeeze; (4) the squeeze grasp; (5) the hand grasp; (6) the palm grasp; (7) the superior palm grasp; (8) the inferior forefinger grasp; (9) the forefinger grasp; and (10) the superior forefinger grasp. The numbers below the drawings show the ages, in months, at which each of the above stages is likely to appear. (Reproduced by permission of the author and the publisher from H. M. Halverson, "An Experimental Study of Prehension in Infants by Means of Systematic Cinema Records." *Genetic Psychology Monographs*, 1931, 10: 107-286.)

manual prehension by Halverson (1931) merit particular note. Shirley's very intensive and complete account of the motor development of the babies whom she studied does not present anything that can be called a scale for quantitative measurement. In 1927 Cunningham devised a tentative series of motor tests for children between the ages of twelve and thirty-six months, arranged in the form of a year scale with from five to nine test items at each six-month interval. The chief interest in this scale—if it may be so called—arises from the fact that all the items used involve gross bodily activity rather than fine movements of the hand and fingers and relatively few of the tasks set are of the conventional type included in most scales of this kind. Cunningham reports correlations with the Kuhlmann-Binet of approximately $+ .50$ for single age groups and a retest correlation of $+ .58$ for twenty-three cases examined at the ages of twelve and eighteen months. Taken at their face value, these figures are as high as or higher than those reported by Bayley for her more carefully developed scale. It would be well worth while to try working out two separate motor scales for young children, one of which would include only items having to do with the growth of control in the large muscles of the trunk and limbs, the other only those involving movements of the hand and fingers.

The correlations of the motor scores made before the age of six months with those earned after the age of twelve months by Bayley's group were but slightly better than chance. In an unpublished master's thesis done at the University of Minnesota, Patricia Neilon describes the development and behavior of those members of Shirley's group whom she was able to relocate at the age of seventeen years. Only descriptive accounts of their motor skill based upon interviews with the children and their parents are presented since no formal motor tests were given. However, the data are presumably free from halo effect, since Neilon was careful not to acquaint herself either with the pseudonyms used by Shirley to denote the individual children or with the individual reports included in Shirley's monograph until after she had completed her study. Such facts as are given tend to support Shirley's belief in an innately determined motor talent, but the cases are too few and the evidence is too uncertain to justify much confidence in the results.

MOTOR DEVELOPMENT IN LATER CHILDHOOD AND ADOLESCENCE

The Oseretzky (1925) scale of motor tests has an age range from four to fifteen years. In form it is modeled after the Stanford-Binet with six items at each yearly age level, each of which counts for two months

of "motor age." The method of administering and scoring is similar to that of the Stanford, and motor ages and motor quotients are calculated in like manner. The scale attempts to cover all the important aspects of motor behavior using both the large and the small muscles. It includes synchronization of movements of different parts of the body and even facial mimicry. It has been translated into several languages and has attracted a fair amount of attention in the United States, especially during recent years. There is insufficient evidence, however, to establish either the dependability of the scores as measures of the level of motor maturity of children at the time of testing or their predictive value for later motor growth. More research of an objective nature is needed.

Other measuring devices, both general and specific, for use at these ages have been worked out and more or less completely standardized. Whipple's *Manual of mental and physical tests* (1919, 1921) describes many of the classical types of these tests, and in spite of its early date is still a valuable source book. The American Physical Education Association has devised a number of tests of motor fitness and motor achievements, involving such items as the hurdle jump, the broad jump, various tests of throwing, balancing, kicking and the like. There are also achievement tests for the various sports such as swimming, archery, basketball, or volleyball. Allied to these in some ways are the paper-and-pencil tests of sports information, health knowledge, and safety rules.

Among the tests of general motor development, the scales by McCloy (1934) for elementary school children and the Brace Motor Ability Tests (1927) for children in the upper elementary grades and the high school are among the best known. The latter were included by Espenschade (1940) in her illuminating study of the changes in motor performances of boys and girls during adolescence, some of the results of which are shown in Figure 28.

A glance at Figure 28 shows that for each of the motor skills studied by Espenschade, the scores made by the boys markedly exceed those made by the girls, and that in most cases these differences steadily increase between the ages of 13.0 and 16.4 years. The reason for this increasing difference is largely due to the consistent gain made by the boys, while the scores earned by the girls show but small increase or, in the case of the broad jump, actually decline with advancing age. Espenschade ascribes this to lack of interest in motor prowess on the part of the girls as they move into adolescence, quite as much as to physical differences in muscular strength. It is probable, however, that multiple factors rather than a single isolated cause are responsible. Cultural ideas of appropriate activities for the two sexes undoubtedly operate to strengthen the tendency to cease competing when competition becomes more diffi-

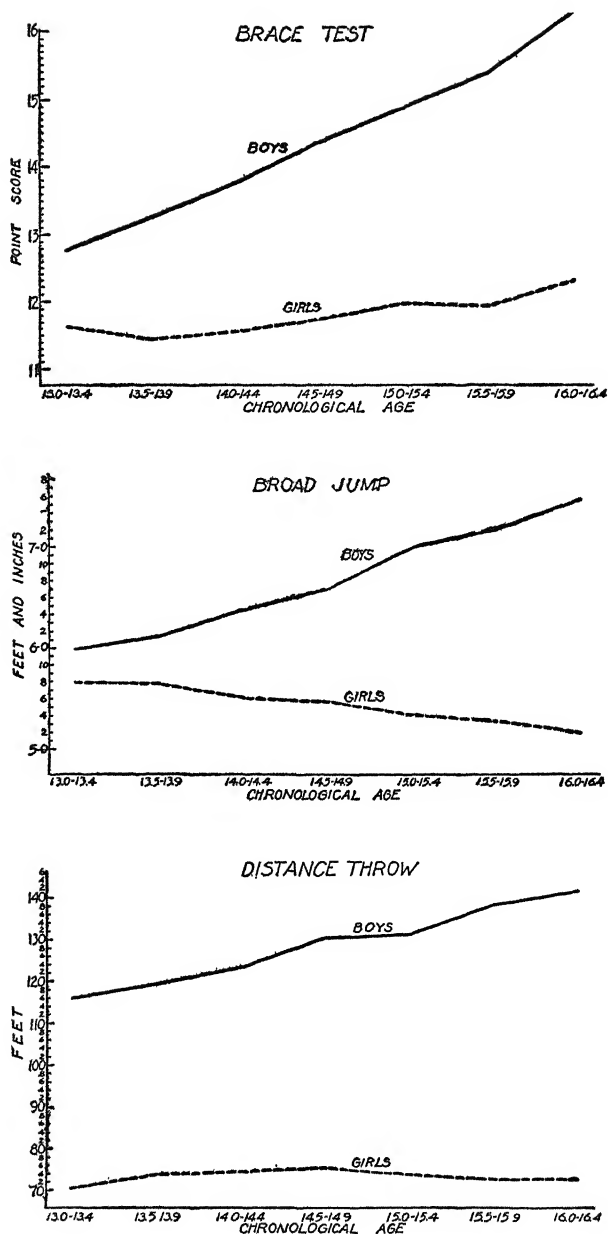


FIG. 28. MOTOR PERFORMANCES IN ADOLESCENCE. (After Espenschade in "Motor performance in adolescence: including the study of relationships with measures of physical growth and maturity." *Monographs of the Society for Research in Child Development*, 1940, 5, 126 pp.)

cult, just as success is a spur to increased effort and to continued practice. Whatever the causes may be, the effect is unmistakable.

In addition to the tests of gross motor ability, two tests of the speed of fine hand-movements were made. The first of these was the ordinary test of reaction time, which consists of an electrical measurement of the quickness with which the subject can respond to an auditory stimulus by lifting his finger from a button. The second was a measurement of the time required to move a peg from one hole to another. This likewise was electrically recorded. None of the correlations between fine and gross activities were large enough to be dependably greater than chance except those for the Brace tests, where the average for several determinations was in the neighborhood of $+.35$. Although not high, an r of this magnitude (for eighty cases) exceeds the value needed to reach the 1 per cent level of confidence. Espenschade suggests that this indicates that the Brace test is more nearly a measure of general motor ability than are the others. Inasmuch as the test includes a total of twenty items, while the others involve single performances only, this may well be true.

MOTOR TESTS FOR COLLEGE STUDENTS AND FOR UNSELECTED ADULTS

The tests of motor ability used at the college level are generally designed either to assess general physical fitness or to determine aptitude for some particular sport. In content and method they do not differ materially from those used in high school except that somewhat higher standards of physical strength are imposed. Frequently they have been developed to fit the requirements of a particular experiment rather than for general use. Like those used at the younger ages, the intercorrelations of the separate items are usually not high.

Motor tests for use in occupational or vocational guidance or the selection of candidates for industrial positions have an important place at these ages. These tests will be considered in a later chapter.

THE MEASUREMENT OF MUSCULAR STRENGTH

Strength is commonly measured by the use of a dynamometer, an instrument that records the force of muscular exertion. There are a number of different types adapted to measuring the strength of different muscle groups. One of the most commonly used is the hand or squeeze dynamometer for measuring strength of grip. The Martin Strength Tests (1921) are designed to measure the strength of the major muscle

groups of the entire body. They consist of a system of straps attached to steel springs with dials on which the amount of force exerted against them is registered in the same way as on an ordinary steelyard. The straps are adjusted to the subject's body. After being placed in a specified position with hand or foot in a stirrup connected with the measuring apparatus, he is told to pull or push as strongly as possible. Several trials with interposed rest periods are usually given. The arrangement

of the harness is such as to make it possible to measure the strength of different parts of the body on each side separately. A comparison of the strength of the right and the left sides is sometimes used as a measure of laterality on the supposition that greater use will result in increased strength.

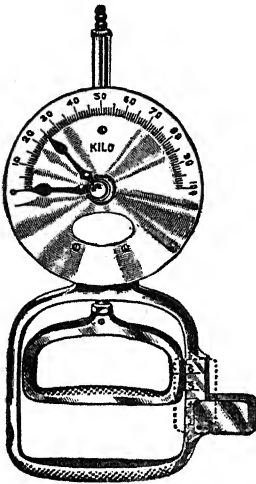


FIG. 29. A HAND DYNAMOMETER. (Courtesy of the C. H. Stoelting Company.)

A special type of dynamometer traces a record on a revolving drum. Such a record is sometimes called a dynamograph. The ordinary dynamometer records the maximum force only; the recording dynamometer makes it possible to study endurance as well as momentary strength.

A number of other methods of testing strength are used, such as lifting weights, "chin-ning," and other athletic stunts that can be measured in terms of duration or number of performances. Since most involve some degree of skill as well as strength, and this tends to confuse the issue, none of these are quite as dependable as the dynamometer tests.

Katz and McLeod (1939) devised a special type of dynamometer for measuring the strength of the jaws in biting. They were particularly interested in studying what they called the "mandible principle," that is, in comparing the sensation of strain when a given result is achieved by means of the opposed action of two parts of the body in a pincerlike movement with that experienced when only a single part is involved or when two parts cooperate by moving in the same manner and in the same direction. They found that when subjects were told to lift a certain weight and then to squeeze a hand dynamometer to a point where the same effort was required as in lifting the weight, the dynamometer estimate was always many times greater than the amount of the weight used as the standard. The difference was even more striking when the jaw dynamometer was used.

TESTS OF LATERAL DOMINANCE

That one hand is commonly used in preference to the other, especially when acts of skill are to be performed, is known to everyone. Most people prefer to use the right hand, but an appreciable number prefer the left. Not so widely known is the fact that foot preferences also exist and that one eye usually takes the lead in aiming and sighting. The preferred side is commonly called the "dominant side," and the measurement of the extent of preference is called a measure of hand dominance, eye dominance, or foot dominance, as the case may be.

The study of lateral dominance was given tremendous impetus when Samuel Orton (1927) and Lee Edward Travis (1931) put forth the theory that a major cause of stuttering is to be found in incomplete dominance of one cerebral hemisphere over the other. That the control of the speech mechanisms is vested in a single cerebral hemisphere had long been known to neurologists. The speech center, known from its discoverer as the *region of Broca*, is located on the left side of the brain in persons who are predominantly right-handed; on the right side in those who are left-handed. Hand preference thus indicates which side of the brain exerts control over the speech mechanisms. The hypothesis that such a disturbance of the pattern and rhythm of speech as occurs in stuttering might arise from interference between the speech centers in the two hemispheres when neither side is dominant over the other was a natural outgrowth of expanding neurological information, and the idea gained in strength and popularity when a number of studies appeared to show that various anomalies of lateral dominance were more frequent among stutterers than among those with normal speech. These anomalies include ambidexterity, right-handedness combined with left-eyedness, or vice versa (mixed dominance), cases in which the right hand is preferred for certain operations, the left for others, or instances in which "native" handedness has been changed by forcing the child at an early age to use the right hand for practically all manual performances when his natural preference was for the left.

A review of the evidence on this much-disputed problem would be outside the scope of this book. We shall note only that the theory that incomplete or changed lateral dominance is the sole or even the major cause of stuttering is less widely accepted today than it was a decade ago, though there is evidence that it may be a contributory factor in some cases. The associated theory that help for the stutterer can be had by training him to use the nonpreferred hand, in the belief that by so doing he will gradually restore the cerebral dominance which (it is inferred) has been disrupted by early training or experience, is no

longer accepted by many people. But the entire concept has given rise to much interest in the measurement of lateral dominance and to the development of many tests designed for that purpose.

Tests of hand dominance are many. Some involve single functions only. Among these, a comparison of the rate of tapping with the right and the left hands, either separately or simultaneously, and the difference in the speed of reaction of the two hands are old favorites. Tests of rapid coordination of movement, such as the "three-hole test" in which the subject is required to thrust a stylus into each of three small circular holes in rapid succession so as to make an electrical contact at the bottom, are also widely used. The number of thrusts is registered on an electrical counter. As the diameter of the holes is only slightly larger than that of the stylus, a comparison of the relative efficiency of the two hands in a task of this kind is a more delicate test of dexterity than is afforded by a more mechanical performance such as tapping.

Other single devices sometimes used are instruments for measuring hand steadiness, pursuit meters of various types, needle threading, and mirror drawing. Many people, however, noting that few people are wholly consistent with respect to hand preference for simple activities, prefer tests involving a variety of acts. Some rely on questionnaires filled out by the subject, who is asked to state which hand he commonly uses for a number of specified acts. Several standard lists of such questions have been published. Others prefer actual tests using a variety of short tasks. Usually the subject is kept in ignorance of the purpose of the test in order to avoid self-consciousness and to ensure that his responses will be natural and spontaneous. One of the best of these tests is that by Wendell Johnson (1936, 1937), which includes sixty-four very simple tasks, such as lowering a window shade, tearing a sheet of paper from a pad, taking a crayon from a box, putting a cap on a fountain pen, and others of a similar nature. Johnson reports very high self-correlations for this test, which is easily administered and requires only such apparatus as is found in practically every home or office.

Tests of eye dominance are usually made with a simple device known as a manoptor or manoptoscope (sometimes also called a V-scope). A simple but serviceable model can easily be constructed from lightweight cardboard according to specifications given in Figure 30. The devices can also be obtained commercially. A rough test of eye dominance can be had by holding the index finger at arm's length and covering each eye successively. When the nondominant eye is covered, no change in the apparent position of the finger will occur, but when the dominant eye is covered, so as to force sighting with the eye not commonly used for

that purpose, the finger will be seen at a different angle and so will appear to move to one side of its former position.

Foot dominance is not as easily measured since there are few activities for which only a single foot is used. Even in such operations as turning a crank by means of a foot pedal, it is not always clear whether the foot that operates the pedal or the one that supports the weight of the body while the work is going on is filling the dominant role. Although a few attempts at standardizing tests for foot dominance have been made, they are not very dependable. Tests of the comparative reaction time of the two feet and of their relative speed of tapping are perhaps as satisfactory as any that have been used up to the present time.

THE CORRELATION OF MOTOR SKILL WITH OTHER ABILITIES

The tests used for the measurement of what is called "intelligence" and those used for the measurement of what is called "motor ability" during infancy and very early childhood have so much in common with each other³ that it is not surprising to find that the correlation between the two presumably different types of measures is generally almost, if not quite, as high as the self-correlations of either when the testings are separated by a few months of time. Among older children the relationship between tests of intelligence and those of motor skill or muscular strength is still positive, as a rule, but considerably lower. Examination of the scatter diagrams will usually show that much of the obtained correlation is attributable to the lower levels of intelligence. Backward children and adults are typically awkward in their movements; their gait is frequently shambling and their step heavy, lacking in resilience and grace. Their hands are clumsy and although many of them can be taught such manual skills as basketry, weaving, or lacemaking, their rate of learning is slow. Among children or adults of normal intelligence, however, the relationship between mental and motor abilities, although still positive when large groups are considered, is very low. The correlation is slightly higher when measures of motor strength are substituted for measures of speed or coordination, but it is questionable whether the differences between bright and dull are due wholly to actual differences in strength of muscle or are attributable in large part to more effective use of such strength as is possessed on the part of the brighter subjects in the performance of the tasks set.

³ For example, some of the items used in Bayley's scale for measuring the motor development of infants are identical with those included in her First Year Mental Scale.

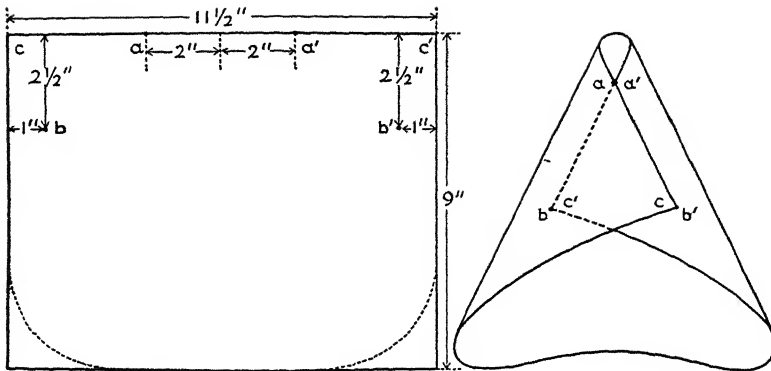


FIG. 30. SPECIFICATIONS FOR CONSTRUCTING A MANOPTOSCOPE FOR DETERMINING EYE DOMINANCE. (Reproduced by permission of the D. Appleton-Century-Crofts Company from *Developmental Psychology* by Florence L. Goodenough.)

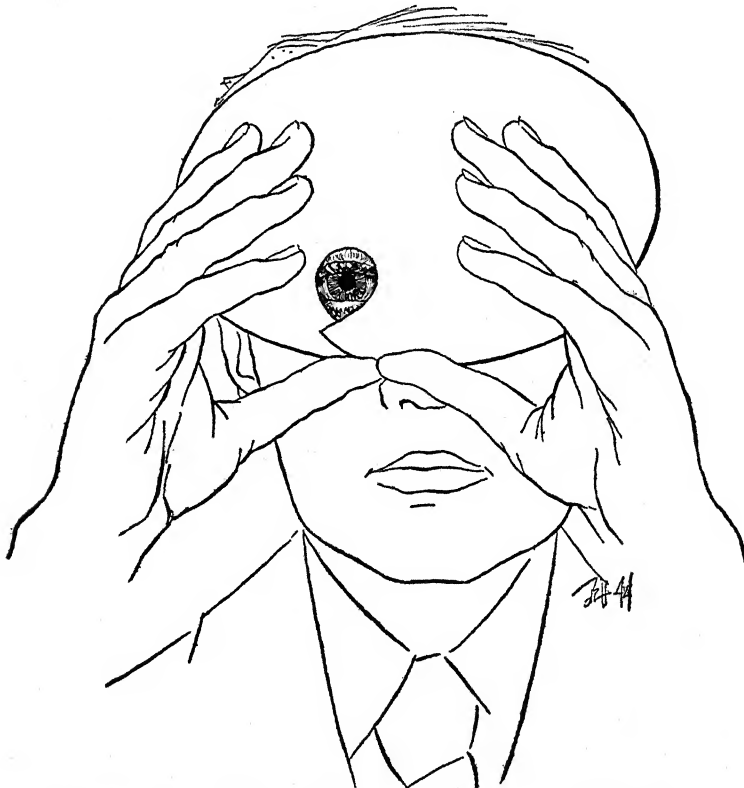


FIG. 31. ILLUSTRATING THE MANNER OF USING THE MANOPTOSCOPE DESCRIBED IN FIG. 30. (Reproduced by permission of the D. Appleton-Century-Crofts Company from *Developmental Psychology* by Florence L. Goodenough.)

At all ages, some relationship between body form and motor skills has been noted. Both Bayley and Shirley found that children with long legs and short bodies tend to walk alone at an earlier age than do those of the opposite physical type. All other things being equal, long-fingered persons have an advantage over those with short fingers in performing certain acts of manual skill.

The relationship of motor skill to socioeconomic status is generally somewhat greater for fine manual operations than for gross bodily skills. However, when the group includes a wide range of nutritional and health conditions this tendency may be obscured or even reversed, if, as is often the case, health conditions are related to socioeconomic level.

The Measurement of Interests and Attitudes

PURPOSES AND AIMS

In the study of interests and attitudes, two rather distinct purposes can be noted. In the one case the facts obtained are looked upon as samples of a larger body of data of the same kind. As in other sampling procedures, the purpose of the sample is to provide an estimate of the character of the universe from which it is drawn. In the second instance, the facts secured have but little interest in themselves. They are regarded not primarily as samples but as signs by means of which some other characteristic of the individual who exhibits them may be inferred with more or less exactness. In the first instance the criterion is intrinsic; the only questions that arise have to do with the representativeness and the adequacy of the sample and the factual accuracy of the data secured. In the second case, the criterion is extrinsic and is of major importance. In this chapter, studies of the first type only will be considered. Those belonging to the second group will be discussed in connection with the topics to which they refer.

METHODS OF STUDYING INTERESTS

In any description of the characteristics of other people, some mention of the things which they seem to enjoy, of their specific likes and dislikes, is almost certain to be included. These informal notes and observations are often given more systematic form by means of check lists, questionnaires, or rating scales.

Choice tests are frequently used, particularly with young children for whom a verbal expression of interests is not likely to be dependable unless given some direction by means of a simplified situation in which the number of alternatives is reduced in conformity with the child's level of mental development. The method consists in asking the child which of two or more objects or activities he prefers. Particularly with very young children, pictures of concrete objects not easily accessible may be used. Differences in the interests of groups separated on the

basis of such factors as age or sex, home background, or other circumstances are frequently studied in this way. In some instances the various choices are treated as discrete units without attempt at combination. Sometimes they are grouped according to their overt character, such as active versus quiet games, mechanical toys versus dolls or "cuddly" animals, and so on.

When the subjects are able to read, the check list is a common method of studying interests. A list of activities, occupations, or similar items is provided, and the subject is asked to check those he prefers. The manner of checking varies. The subject may be asked to place a double check before items particularly enjoyed, a single check before those moderately enjoyed. He may be asked to specify the one (or more than one) most liked. Instead of rating interests as such, he may be asked to check those activities in which he has engaged one or more times during the preceding week or month. In this case, participation is taken as a measure of interest. As in the experiments involving choice, the items may be considered separately or grouped according to kind, but in either case they are taken at their face value, not as signs of something else.

When a number of items are grouped to form a single category, however, an outside criterion is sometimes desirable to serve as a guide to the grouping.¹ Such a criterion may take the form of the pooled judgment of a number of observers or of some objective measurement of the character of the items. Games or activities other than games may be grouped according to some more or less objectively determined characteristic they possess in common. For example, reading material may be grouped into fiction, poetry, biography, current affairs, and so on. The study of preferences is closely related to the study of interests, but is somewhat more narrowly defined since a person may prefer a single one among a number of things in which some interest is felt.

Closely associated with the study of interests is the "values" test devised by Woodruff (1941). Decisions of major importance must sometimes be made by all of us. Working upon the principle that in the nature of these major choices is to be found the key to the deep-lying interests, the basic *values* that each of us assigns to the different goals for which men strive, Woodruff devised a series of short paragraphs in which various contrasted situations designed to resemble some of the

¹ In a sense, of course, the judgment of the person who makes the grouping may always be regarded as an "outside criterion." The reference here, however, is to those instances in which the basis for grouping is not so objective but that some disagreement among judges is likely to occur, which makes a criterion other than the opinion of a single judge either necessary or highly desirable.

important choices that have to be made in real life were described. None of the alternatives was wholly ideal. All combined some one feature that is generally thought of as desirable with certain drawbacks. The "values" selected by Woodruff were as follows: wealth, society, political power, social service, pleasant home life, comfort and ease of living, religion, occupational and financial security, personal development, adventure and excitement, friendships, and intellectual stimulation and activity. Each of these was stressed in two of the following settings: choice of a locality in which to live and work, choice of a college fraternity or sorority, and choice of a vocation. The vocations were not named but merely described in terms of the opportunities afforded and the disadvantages likely to be present in each.

Woodruff's study is exceedingly suggestive as far as its general purpose and approach are concerned. His statistical procedures leave much to be desired. It is to be hoped that in the future someone who is better versed in the techniques of test construction will subject the idea to more refined treatment.

A particularly valuable feature of Woodruff's method lies in the fact that it suggests a way of handling the difficult problem of measuring the strength of interests. Both the general pattern of interest and the breadth of interests are by comparison easy to determine. But there is a question of even greater importance that most of the conventional methods of testing leave untouched. That is the question at which Woodruff is fumbling. Not only the kind of values, the character of the goals for which an individual strives but the energy and single-heartedness with which he pursues them, the sacrifices that he is ready to make in order to attain them are of equal or even greater importance. It is, of course, possible that a verbal description may be so far removed from the actual situation that it cannot afford any useful indication of what is likely to occur in the face of hardships really experienced and rewards partially attained or more realistically foreseen. But if a test on the order of Woodruff's were to be devised in which both the goals to be pursued and the obstacles to be overcome were scaled by a method similar to that proposed by Guttman (1944), it is possible that a measure of much significance might be developed.

The Allport-Vernon Scale of Values (1931) has a longer history and has been more widely used. It is based upon a classification originally proposed by Spranger of which the third edition was published in 1922. Because of the nature of the "values" studied, the Allport-Vernon scale has been rather widely used in vocational guidance. Unlike the Woodruff test, this scale makes no attempt to measure strength of interest-values but merely provides a means of arranging six classes of interests in order

of preference: (1) theoretical interests, characteristic of the scientist or the philosopher; (2) economic interests, characteristic of the man of business; (3) aesthetic interests, characteristic of the poet, the artist, or the musician; (4) social interests, or interest in people, characteristic of the social worker, the salesman, or the philanthropist; (5) political interests, characteristic of the politician or the diplomat; (6) religious interests, characteristic of the clergyman or the mystic. Retest correlations on this scale are fairly high for most of the categories when used with college students, but are low for the measure of social values, which in general appear to be very poorly represented. It is apparent that the general approach is very different from that of the Woodruff scale. It is a test of preferential interests, rather than a test of values in the ordinary sense of the word.

PRACTICAL APPLICATIONS OF THE MEASUREMENT OF INTERESTS

Generally speaking, samples of interests are secured for the sake of obtaining information regarding scientific problems such as sex differences, changes in interests with age, or differences in the interests of specified groups. There are, however, many practical situations on which measurements of interest will throw light. For example, a superintendent of schools may wish to know which teachers are most effective in arousing the interest of their students in the subjects of the school curriculum. A simple experiment of the check list kind will be suggestive. Children in the various classes are asked to mark the subjects that they like very much with two *L*'s (LL), those that they like fairly well with one *L*, those that they neither like nor dislike with an *N*, those that they rather dislike with a *D* and those that they dislike very much with two *D*'s (DD). Tabulation of the results by classes should provide valuable information as to the attitudes toward classroom work that the different teachers are able to inspire. Teachers who have been outstandingly successful in arousing interest in particular subjects might be asked to describe or to demonstrate their methods to the remainder of the group. The effect of certain outside activities upon schoolwork may depend largely or wholly upon their influence in augmenting or decreasing interest on the part of the students. Objective study of such questions is worth far more than unverified opinion.

METHODS OF STUDYING ATTITUDES

An attitude, as defined by Warren is "a stabilized set or disposition." Used specifically, with reference to a particular object, activity, or

situation, this definition implies a somewhat stable tendency to respond in a certain way in accordance with the individualized meaning that the situation has taken on for the subject in question. Attitudes vary along several coordinates. They differ in *direction* accordingly as they are favorable or unfavorable. They differ in *strength* or *intensity*. They differ in *breadth* accordingly as they extend to all aspects of the situation in question or only to certain specific aspects of it. They differ in *fixity* or *fluidity*. Some people are like weathercocks. Their opinions may be strong for the moment, but they veer about with every changing wind. Others are as obstinate as the proverbial mule. Once an attitude has been established, neither the opinions of others nor objective evidence seems able to move them from their stand. This inflexibility of attitude is not the same thing as strength. It may actually arise from an unconfessed awareness of weakness.

Most devices for measuring attitudes have to do with a single situation or institution and are limited to the development of quantitative scales for indicating the direction and the intensity of the attitude at the time of testing. A very high score thus means a highly favorable (or unfavorable) attitude toward the object in question; a very low score indicates a strong feeling in the opposite direction, with the point of indifference somewhere between the two.

As is true in most fields, modern methods of studying attitudes were preceded by descriptions and judgments which later on were frequently given uniformity of expression by means of questions and rating scales. For the modern methods of constructing attitude scales we are chiefly indebted to Thurstone. The procedure most commonly used is known as *the method of equal-appearing intervals*. In the construction of a scale by this method, the first step is to secure as wide an expression of opinion concerning the point at issue as possible. Suppose it is desired to construct a scale for measuring attitudes toward socialism. A large number of people with a wide range of political affiliations and of varying socioeconomic status and educational background are asked to write down informally what they believe to be true about socialism as a form of government and a way of life. From the mass of data secured in this way a large number of statements are chosen according to the following criteria: (1) the series as a whole should comprise a range of opinions from the most to the least favorable; (2) each statement should be short, pithy, permitting (as far as can be judged) only one interpretation; (3) each statement should embody a single idea only (no double-headers—no “ands” or “buts”—should be included); (4) each statement should be put in such a form that it can be either accepted or rejected; (5) each one should imply something of importance with respect to the institu-

tion in question—in this case, socialism; (6) as far as the experimenter can judge, they should be fairly evenly spaced along the continuum from most to least favorable, and should include some that are near the mid-point of indifference; (7) the number chosen should be definitely greater than the number that it is proposed to retain in the scale.²

Each of the statements to be tested is then typed on a separate card. Each member of a second group of subjects, numbering if possible not fewer than two hundred, is then asked to arrange the cards into a specified number of piles, usually eleven,³ on the basis of their favorableness or unfavorableness toward the topic to be considered. The judges are asked to try to make the arrangement in such a way that the differences in the favorableness of the expressed attitudes between the first and the second pile will be equal to that between the second and the third, and so on to the last. This apparent equality of the step intervals has given the method its name—that of *equal-appearing* (rather than “equal”) intervals.

The judges are told to make as many rearrangements of the cards as seem desirable until they are satisfied that they can do no better. When all the arrangements have been tabulated and pooled, the data are examined for the final selection of items to be included in the finished scale. The aim is to secure a series of statements (1) that are fairly evenly spaced, (2) that cover very nearly the total range from one extreme to the other, and (3) about the position of which there is as little disagreement among the judges as possible. In deciding which items to include, the median position given to each by the entire group is calculated.⁴ A measure of variability is then found for each of the items.

² Most of the attitude scales that have been standardized contain from fifteen to forty items. It is not necessary to decide in advance how many are to be retained, but a sufficient number should be tried out to provide for the eliminations that will almost certainly be found desirable.

³ An odd number of piles is desirable because it permits each judge to begin by classifying the cards into three groups—definitely favorable statements, definitely unfavorable statements, and statements that imply uncertainty or indifference. The end piles are then subdivided into more or less favorable statements, and the process is repeated until the requisite number of piles has been reached.

⁴ The reason for using the median rather than the mean position of the item is twofold. A few of the judges may hold views that differ very widely from those of the majority, either because of lack of acquaintance with the topic or as a result of some unusual personal experience. To allow these rare cases to exercise their full effect upon the scale values would probably be a mistake. The second and more important reason has to do with the “end effect.” Deviations in judgment from the mid-values of the central piles can take place in either direction, but in the two end piles the only possible deflection is toward the mean of the group. It is not possible to correct for this completely, but as the median will be less affected by it than the mean will be, the former is the measure generally preferred. It should be noted that in making their arrangements the subjects should be instructed that it is not necessary to place the same number of cards in each pile.

Different investigators have used different ways of computing variability. As long as the same measure is used throughout, the choice is probably not of great importance. Kelley (1923) has shown that the 10-90 percentile range which he designates by *D* is the most dependable indication of the stability of the median. This is easily computed and is probably to be preferred to *Q* (one half the range of the middle 50 per cent), which has more frequently been used.

Items which show large variability in judgments are eliminated on the assumption that they probably convey different meanings to different persons. When two or more items are found to have been assigned almost identical median positions, the one selected is the one that shows least variance in judgment or has to do with an aspect of the topic not well covered by other items in the series. The process of elimination is continued until a series has been chosen in which the median values are fairly evenly spaced at intervals which should, for best results, not be coarser than 0.5 of the distance from one pile to the next and with no item included for which the dispersion of the ratings is large. It should be noted that in the calculating of medians, as well as in the making of judgments, each pile should be thought of as representing a range of values and not as a point at which all are concentrated. The lowest pile in the series thus represents a range of 0-1, the second has a range of 1-2, and so on. The mid-values are then taken as 0.5, 1.5, and so on. The value finally assigned to each of the items is the median of all the ratings given to it.

In using the scale, the subject is told to place a check before each item with which he agrees. His score is the median value of the items checked.

Of the number of other methods of constructing attitude scales proposed, only two have been sufficiently used to require consideration here. In an attempt to save time and labor by doing away with the use of a judging group, Likert (1932) proposed that subjects be asked to express the *extent* of their agreement with each of the items making up an attitude scale, instead of merely indicating agreement or disagreement in an all-or-none fashion. The terminology of the investigators who have utilized Likert's method varies to some extent, but in general it conforms to the following scheme: *strongly agree*, *agree*, *undecided*, *disagree*, *strongly disagree*. Arbitrary weights from 1 to 5 are assigned to these ratings in such manner that either strong agreement with a statement that favors the institution in question or strong disagreement with one that is unfavorable to it is given a weighting of 5, while those at the opposite extreme are assigned a weighting of 1. Weightings from 2 to 4 are given to the intermediate judgments. The

score of any individual is the sum of his ratings on all the items. Likert's method has appealed to many persons because of its apparent simplicity and because it enables the investigator to start directly with the group of subjects in whom he is interested without the necessity of preliminary standardization of his instrument beyond some form of item analysis used as a basis for the selection of the statements to be included in the scale.

The relative merits of the two methods just described have never been adequately tested. Certain comparisons, however, can be made. Those who have preferred Likert's method have commonly selected items of rather definitely favorable or unfavorable import. Neutral feelings are then indicated by rating the attitude toward that item as "undecided." Thurstone, on the other hand, has attempted to include in each of his scales one or more statements that in themselves indicate a neutral attitude. Whether or not there is a psychological difference between agreeing with a neutral statement and unwillingness to take a decisive attitude toward statements expressed in definitely favorable or unfavorable terms is not known, but the findings of Rundquist and Sletto (see page 407) suggest that the matter is worth investigating. Advocates of the Likert technique have commonly included a greater number of items in their scales than have those utilizing the method of equal-appearing intervals. This means that more time will be needed for scoring the papers, a fact that to some extent does away with the advantages of eliminating the judgment group. Likert's method assigns equal weight to all the items although the chances that all indicate equally favorable or unfavorable attitudes are slight. While it is theoretically possible that inequalities of this kind may be balanced by the use of the differential ratings (a rating of 5 on a moderately favorable item corresponding to one of 4 on one that is extremely favorable), the probability that this will occur has never been demonstrated. Nevertheless, it has been repeatedly shown that if the items are well chosen and properly worded, the self-correlations obtained for scales developed by the Likert method are high enough to be satisfactory for most of the purposes for which such scales are likely to be used. Although a number of investigators have attempted to compare the two methods with respect to the internal consistency of the results obtained by their use, most if not all of these studies may be criticized on the ground of technical flaws that invalidate their findings. More adequate study of the question is needed.

In addition to the method of equal-appearing intervals, Thurstone has worked out a technique for scaling preferences by the use of paired comparisons. Every item in a series is paired with each of the others in

turn. In each case the subject is asked to indicate which of the two he prefers. The procedure is laborious, but it has the advantage of yielding for each of the items a standard score which has a relatively small error of estimate. An additional feature of much importance is the fact that the method permits a comparison of the variabilities of different groups or of the same group under different circumstances. In his original report, Thurstone (1928) applied this procedure to a study of nationality preferences among college students. Twenty-one nationalities were compared. As Thurstone points out, the range of the standard scores is a measure of the amount of prejudice within the group in question. The method thus becomes a valuable means of studying changes in attitudes. If, for example, a group of students was tested before and after a course of study in international relations which stressed the contributions of each nation to world progress, some evidence as to the effect of the training upon racial-national prejudice could be had. In like manner, the effect of propaganda with respect to some particular group could be determined by noting whether or not its standard score position underwent a significant change. The method is thus especially adapted to studying the effect of motion pictures, newspaper campaigns, and so on. It is chiefly a method for measuring group differences rather than for the study of individuals, and for making broad comparisons among discrete classes rather than for more specialized study of attitudes toward a single class or category.

PRACTICAL APPLICATIONS OF THE MEASUREMENT OF ATTITUDES

In addition to the problems just mentioned, a number of other practical uses of attitude measurement may be noted. Inasmuch as attitudes are the springs from which overt behavior rises,⁵ they are unquestionably to be reckoned among the most important areas of psychological measurement. Confidence or distrust, tolerance or intolerance for the ways of others, ruthless competition or mutual helpfulness—these are factors that make for war or sustain peace, that consolidate a community or tear it apart, and that largely determine family harmony and individual happiness. The source of particular attitudes is not always easy to determine. They spring from individual experience, from rumors and reports, from newspapers, magazines, and books. Attitudes

⁵ This does not mean that attitudes always give rise to overt actions. They may be repressed, or behavior which is not their normal outcome may be intentionally substituted as when the well-bred hostess masks under a reassuring smile, her distress over the breakage of a treasured piece of china by her guest, and asserts that the accident is of little consequence.

are strengthened by concepts of self-interest, for all of us tend to believe what we wish to believe.

It is largely because personality finds one of its main channels of expression in the attitudes which an individual assumes toward other persons, specified activities, and social institutions that attitude measurement has become such an important tool of psychological research. The potential usefulness of the method has by no means been completely realized. For example, in the clinical study of an individual not only is it important to know his attitudes toward specified subjects; it is even more important to ascertain the general trend of these attitudes. Tolerance or intolerance, hopefulness or pessimism, cynicism or approval, as well as a host of other tendencies, appear to become more or less generalized reactions in the majority of persons as age advances. Factorial analysis of the items included in a wide range of attitude scales might provide the data for the development of scales for the measurement of these more general attitudinal characteristics that would be of much value for the clinical worker. A few attempts along this line have already been made, but more are needed.

The rigidity or flexibility of individual attitudes is another problem which merits more study than it has received. A number of excellent experimental studies on the effectiveness of various methods of modifying attitudes have been made, and although all such studies have shown that there is considerable individual variation in the effect upon different subjects of the procedure used, little if any attempt has been made to relate these differences to other characteristics of the persons concerned. It is generally believed that young persons are more flexible in their attitudes than older ones, but not much in the way of concrete evidence has been adduced in support of the belief. Certainly the rule is one that has many exceptions.

SOURCES OF ERROR IN THE MEASUREMENT OF INTERESTS AND ATTITUDES

Paper-and-pencil tests of the kind described in this chapter obviously do not always yield dependable information about the actual behavior of the respondents in the situations described. Strictly speaking, all that they provide is a series of claims which may or may not conform to the behavior under consideration. Many people who profess highly favorable attitudes toward certain minority groups, such as the Negro, will nevertheless consult a White physician of relatively mediocre training and ability in preference to a Negro of far greater competence. Certain interests and attitudes meet with social approval; others are looked upon

with less favor. The natural human tendency to "put one's best foot forward" gives to items that fall in the first class some advantage over those in the second class as far as the likelihood that they will be checked as "agreed with" is concerned.

In the construction of attitude scales it is therefore important to include as few items as possible that impinge strongly upon well-established social conventions. In the interpretation of the results of attitude scales, whether or not they contain such items, the importance of regarding the statements as claims which may or may not justify a prediction of behavior that corresponds to the claims can hardly be overestimated. It should be noted, however, that the possible importance of a given claim does not rest entirely upon its factual accuracy. Claims are facts in their own right, but their meaning does not always conform to their face value, nor does a given claim necessarily have the same significance for all who make it. Here again a factorial analysis of patterns of response might throw much light upon individual personality.

One of the main difficulties confronting the maker of attitude scales is the question of the limits by which an attitude is bounded. The common assumption that these limits are sufficiently described by naming the object to which the attitude refers is convenient but inaccurate. What, for example, of a scale (of which a number have been devised) for measuring "attitudes toward the Negro"? As verbally expressed, attitudes may be in terms of beliefs, or of statements regarding personal behavior, and in many of the scales now available, both are included. Consider the following:

"No distinction should be made between Negroes and Whites in respect to occupational opportunities."

"If both were equally competent, I would send my child to a class taught by a Negro teacher as readily as to one taught by a White teacher."

Theoretically, all who agree with the first statement should also agree with the second, but actually this is not the case. The first expresses an attitude having to do with an abstract principle of social justice, where the Negro is more of a symbol than a concrete race of flesh-and-blood individuals. The second brings the principle down to earth.

Or consider these:

"Taken as a class, Negroes are likely to have better singing voices than Whites."

"If given equal opportunity, most Negroes would show administrative ability equal to or greater than that of Whites."

Racial prejudice plays but small part in determining responses to the first of these questions, but it is likely to prohibit a fair consideration of the second. In general, attitudes are much less clearly delimited by

the name given to a scale than its authors have supposed. More specific terminology with corresponding restrictions upon the character of the items to be considered would result in greater homogeneity of scale meaning and corresponding reduction of erroneous or biased interpretation.

A major problem in the practical use of attitude scales is the lack of dependable evidence as to the usefulness of these scales for the prediction of actual behavior. It is rather surprising to find that so few careful studies of this highly important topic have been made. Dozens of published scales deal with almost every conceivable type of "attitude." There are scores of reports on scale construction and on the application of particular scales to various groups of subjects. But unless the claims with which these scales deal can be used as predictive measures from which some inkling of actual behavior can be had, their study is more of an academic exercise than a useful tool for genuine scientific research or for the practical guidance of individuals. It might not be a bad idea to put a ban on the construction of further scales of this type until more has been learned about the significance that may justifiably be attached to those now available.

THE PUBLIC OPINION POLL

The measurement of public opinion is an outgrowth of the straw vote. For the details of the methods commonly used at the present time we are chiefly indebted to George Gallup, whose successful prediction of the outcome of the presidential election of 1936 gave dramatic emphasis to the fact that the size of a sample cannot compensate for bias in its composition. Hadley Cantril and his associates in the Office of Public Opinion Research at Princeton University in their valuable book *Gauging public opinion* (1944) have described various modifications of the original procedures together with tests used for evaluating them.

As most people are aware, the public opinion poll is a sampling device which is not fundamentally different from other sampling procedures. As commonly used, trained workers are employed for the purpose of (a) selecting from a given locality a stated number of persons who conform to certain specified conditions required for the total sample, and (b) interviewing these persons to secure their opinions with respect to the questions to be included in the survey. For example, a worker might be instructed to interview twenty persons in each of the following age groups: under twenty, from twenty to thirty-five, between thirty-five and fifty, over fifty.⁶ The members of each group are to include

⁶ Ages are usually estimated rather than asked in order to avoid antagonizing respondents and thus running the risk of securing a biased selection through failing to secure cooperation of many of the cases approached.

either an equal number or a specified proportion of the various socioeconomic groups as determined by occupation, income, or some other reasonably objective criterion. Other qualifications may be specified. Sex is usually taken into account. Sometimes political affiliation, as indicated by the party voted for in the last presidential election, may be noted. The particular terms in which the sample may be designated will vary according to the apparent requirements of the problem, but the general plan of defining the number and characteristics of the part-sample to be interviewed by each worker remains the same. The exact form of the questions to be asked and the amount of supplementary information to be given when necessary are also specified. The workers, in addition to being selected on the basis of their pleasing manner and ease when meeting strangers, are given special training in the art of making their contacts and explaining their purpose in such a way as to gain the confidence and win the cooperation of as many people as possible. Under these circumstances relatively few refusals are reported and the effect of these refusals has been subjected to careful statistical analysis. Cantril (1944, p. 120) reports that in his investigations of this question, the total number of refusals for all causes was about 14 per cent of those approached. Now if these refusals had been equally distributed over the entire sample with reference to such factors as age, sex, or socioeconomic status, the total results would not have been affected at all. Only to the extent that the refusals tend to be more frequent among some groups than others is bias introduced. Such differences, it was found, do exist. Refusals were more frequently met among the old than among the young; among women than among men; among the poor than among the well-to-do; in large cities than in rural districts and small towns. But none of these differences were very large. Cantril's figures indicate that the total extent of bias in the final results that could be attributed to refusals did not exceed 1 per cent in any of the cases in which a special study of this factor was made.

The fact that, when the conventional method of polling is followed, the final choice of subjects for interviewing is left to the person who conducts the interviews obviously introduces the possibility of bias. As was indicated in an earlier paragraph, most of the polling agencies merely give their interviewers general instructions as to the characteristics of the persons whom they are to select for questioning. Each interviewer is told to make his sample conform to a specified set of proportions with respect to age, sex, economic status, and whatever other characteristics are thought to be important. Provided that these conditions are met, the choice of subjects is left entirely in his hands. Inasmuch as interviewers are likely to be chosen from the upper half of the economic and educa-

tional distribution, this practice may well result in systematic bias in respect to the subjects most likely to be chosen.

A method known as "area sampling" has been designed to correct this factor. In this type of sampling the entire country or that part of it which the poll is to cover is first marked off into a large number of areas that are approximately equal as to size of population. As a matter of convenience, these areas are usually made to conform to ordinary political or geographical divisions such as counties or metropolitan districts. One of the standard methods for securing a random sample is used (see Chapter 8), and a sufficient number of these areas is then chosen to make the final sample as large and as varied as is thought necessary for the purposes of the study. Each of the areas thus selected is marked off into units so small that only a very few persons, possibly not more than one or two, who are suitable for interviewing will be found in each. A random sample of these units is then chosen according to the same method used before. The interviewers assigned to the various districts are instructed to get the needed information from *every* person residing within the small units chosen for the final sample. If, for example, an attempt is being made to predict election results, the interviewer must question every qualified voter in each of the selected small units in his territory. If a certain person is not at home when he calls, he must try again and again until it seems reasonably certain that he will not be able to make the contact. In such cases or in instances of persistent refusal, the central organization must be notified.

The method of area sampling automatically corrects for possible interviewer bias, since the choice of subjects is fixed by the conditions imposed.⁷ Area sampling is more costly than the ordinary methods since the task of designing the procedure must be performed with great care and the expense of the actual interviewing is likely to be considerably increased by the repeated calls necessary to secure a 100 per cent sample. However, there is evidence that the added expenditure may be worth while. Likert (1948), in reporting on certain findings obtained by the Survey Research Center of the University of Michigan, states that in a series of five studies designed to test the accuracy of the method, the distribution of such factors as age, amount of schooling, and the proportion of native-born Whites within the samples corresponded very closely to those for the entire nation as given in the census reports. As a rule the

⁷ It should be unnecessary to point out that this statement refers only to the selection of the cases who are to be interviewed. Other types of interviewer bias such as may be introduced through tonal inflection or facial expression when the questions are asked, as a result of the interviewer's own attitude toward the matter, may still occur.

discrepancies did not exceed 1 or 2 per cent. If data from other sources prove to be equally favorable, the method may eventually come to supersede those previously used in spite of its greater cost. Studies made by other universities and by a number of government agencies including the Bureau of the Census have thus far obtained results very similar to those reported by Likert.

In practically all of the modern polls, oral interviews have been used in preference to a mailed questionnaire. People of little education find even a simple form of this kind a rather formidable task which they hesitate to undertake; busy people are likely to drop a questionnaire in the wastebasket with scant ceremony; careless people mislay them; and distrustful ones question their purpose. Not all these difficulties can be overcome, even by the most tactful interviewer, but there is little doubt that a much more nearly representative sample can be obtained through the direct interview than through the mails.⁸ Radio broadcasting of questions has also been tried, but as the poorer classes often do not possess radios, some degree of economic bias is practically certain to result when this method is relied upon. Moreover, since responses must be sent in voluntarily, further bias of an unknown character is highly probable.

Public opinion polls have met with plenty of criticism. That such criticism is not wholly unmerited was strikingly demonstrated in the 1948 presidential election, when the pollsters received so severe a jolt that it will probably take some time for them to recover from it. In the end the result may be beneficial, for it seems quite likely that the success of the 1936 prediction, together with the striking and apparently valid results of other studies by this method,⁹ very possibly gave rise to undue confidence in any and all studies dignified by the name of an "opinion poll." Satisfaction with existing methods led to insufficient attention to the details of sampling. Be that as it may, the 1948 failure should not blind us to the many possibilities of the method as a research tool. On

⁸ A recent study by J. R. Shannon (*Journal of Educational Research*, 1948, 42, 138-141), in which the results of 433 questionnaires sent out by reputable agencies were analyzed, showed that only 65 per cent of those sent out by mail were answered and returned while 88 per cent of those handled by personal interview were successfully filled out.

⁹ As McNemar (1946) has pointed out, very few of the results obtained in public opinion polls are subject to verification from outside sources. A belief or an opinion is, by its very nature, a private affair which begins and ends with the individual. In some instances, however, the polls have dealt primarily with statements of *intention* which can be checked against later action or with *attitudes* of a kind to which certain forms of behavior might reasonably be expected to conform. In the latter case some type of objective test might be used to determine the validity of the opinion poll, but unfortunately this has not often been done.

the contrary, it should lead to more careful examination of the procedures used in order that they may be made more accurate. Basically the method of the opinion poll is sound, but the pitfalls in its application are many. The 1948 fiasco should be looked upon as a challenge, calling for improved polling techniques. It should not lead to an abandonment of the method, though it may well show the need for a good many changes in the manner of its application. McNemar's (1946) suggestion that attitude scales consisting of a variety of questions about each topic be substituted for the single questions commonly used in public opinion polls is worthy of consideration, but a change in the method of questioning would have no effect upon errors resulting from an unbalanced selection of the persons to be questioned, which was obviously the factor chiefly responsible for the failure of the 1948 election poll. In this paper McNemar calls attention to a number of questionable features of the polling method as it is often used. Although some of his points have been questioned, the discussion as a whole is highly stimulating and should be read thoughtfully by all who expect to do serious work in the field of attitude measurement or in the gauging of public opinion.

The Measurement of Personal-Social Characteristics

DEFINITIONS

The term "personality" is variously defined by different people. Warren lists five different definitions. MacKinnon (1944) devotes an entire chapter (48 pages) to an account of the various definitions of the term and the attempts at analyzing the "structure" of personality as conceived by leading psychologists of the present day. Kimball Young, in his *Personality and problems of adjustment* (1940), devotes two chapters totaling 61 pages to a discussion of theories and types of personality. Other writers have been similarly loquacious. I shall not attempt to add to the existing plethora of concepts as to what the word "personality" *should* mean but will merely indicate the limits of the term "personal-social" as it will be used in this chapter. Three kinds of measures will be considered. The first type attempts to give quantitative or semi-quantitative expression to some aspect of the relationship between two or more persons. These will be called *measures of social intercourse* or *social participation*. The second group has reference mainly to the affective life of an individual; to his feelings and emotions as given verbal expression by himself or as judged by others on the basis of his behavior. For want of a better term and because measures of this kind have commonly been so designated in the literature, these will be called measures of *personality*.¹ The third is a mixed group. It includes tests or scales in which some of the items would be classed under the first of the heads named above, others under the second. It also includes tests or measures of such a nature that it is difficult or impossible to say whether the social or the personal aspect predominates. We shall designate this mixed group as *measures of personal-social characteristics* or *personal-social behavior*.

¹ That both are present in some degree in practically all situations goes without saying, but in the scoring of a given test only one aspect may be taken into account.

GENERAL METHODOLOGICAL CONSIDERATIONS

In the study of personal-social behavior, two problems arise that have required but scant attention up to this point. The first is a question of *values*. In the measurement of abilities and skills it has been implicitly assumed that greater ability or a higher degree of skill is desirable. The quantitative measure has been thought to run parallel to the social value commonly ascribed to it.² It has been further assumed that a change in the amount or level of the trait could be expressed as a straight-line function, with the higher levels differing from the lower only in a quantitative manner. But as soon as we pass over into the realm of social interaction or that of the affective life, neither of these concepts can be accepted without examination. We do not question the belief that, granted equal comprehension and retention, speed of silent reading is a goal without formal limits. The more rapid the better, we think. But where lies the optimum degree of dominance as opposed to submission—to consider only one of the personal-social traits for which measurements have been proposed? Are we really dealing here with a straight-line continuum?

Quite as difficult to handle is the second question, which has to do with the practical administration of tests or measurements in this field. This is the question of directed motivation of the subjects. In tests of skill or ability the subjects not only are urged to put forth their best efforts but are also instructed as to the manner in which those efforts should be expended. There is no evasion, no half-truths, nothing that smacks of the confessional. One may say to a subject, "I want to see how well you can remember numbers. Listen carefully and repeat these, just as I say them." But suppose, instead, instructions were as follows: "I want to see how readily you can make acquaintance with strangers. If you were meeting me for the first time, how would you get a conversation started? Show me." Not only is such a task likely to arouse self-consciousness and perhaps some degree of resentment on the part of the subject, which would make his response a poor sample of what he would actually do outside the laboratory, but neither the procedure to be followed nor the goal to be reached is clearly enough defined to permit easy scoring.

The sampling method of studying personal-social behavior does not lend itself well to laboratory practice, but it has been used rather exten-

² This concept has occasionally been questioned. Hollingworth (1942), for example, was of the opinion that at least during childhood an IQ above 180 or thereabouts renders adequate social and personal adjustment exceedingly difficult if not impossible, because of the great difference between the child and his mates in respect to general interests and physical and mental maturity.

sively in the study of child behavior in the schoolroom and on the playground. For the most part, however, the aim of these studies has been the determination of group trends, of changes with age, of sex differences, or of the effect of experimentally introduced conditions upon behavior. In the exploration of such factors a number of special methods of observation and recording have been devised which have proved very useful for the purposes for which they were intended. These procedures, however, have been but little used for the study of individual differences.

In addition to direct observation under informal conditions, with its correlates of rating scales and questionnaires concerning the behavior of others, the methods chiefly employed for the measurement and appraisal of social behavior include formal laboratory experiments of different kinds, paper-and-pencil tests and questionnaires filled out by the subjects themselves, and the so-called "projective methods." The last-named, however, have been more often used for the study of personality characteristics, particularly difficulties of personal and emotional adjustment, than for the study of social relationships as such.

METHODS OF STUDYING SOCIAL PARTICIPATION

Two slightly different methods have been employed for reducing to quantitative form observations made under the unstandardized conditions of everyday life. These are known as *time sampling* and *episode sampling*. In the former, the unit of measurement is a single short period of time, and the measure is the number of such time intervals during which a specified form of behavior was observed. The interval is usually made short enough that behavior will be unlikely to change from one category to another during its course. In episode sampling, the unit of measurement is a discrete form of behavior or episode, such as a quarrel, a temper tantrum, a question, an appeal for help. The score is the number of such episodes observed during a longer period of time, say during daily observations of an hour each, taken over a period of two weeks. The time-sampling method is commonly preferred when the behavior studied can readily be classified into a number of categories that form a roughly continuous series with one or another of the specified levels always present, or in the simultaneous observation of large groups when it would not be possible to note all the episodes separately. The method of episode sampling is likely to be chosen when the events observed are conspicuous enough so that they are not likely to be overlooked in group observation and infrequent enough to make the time-sampling method with individual observation wasteful of time, or when the episodes are

of such a nature that they do not lend themselves readily to further analysis in terms of degree or amount.

Arrington (1939) published a well-annotated bibliography of time-sampling studies made up to that time. As no important change in the general procedure has been made since that date, the list may be considered sufficiently representative from the methodological standpoint. Three studies will be described briefly to illustrate the procedures commonly used.

Parten (1932, 1933) observed the social behavior of nursery school children during a series of sixty one-minute periods, using the time-sampling method. Only one child was observed at a time. In order to avoid bias that might result from circumstantial factors affecting the child's behavior at certain times rather than at others, two precautions were taken. Since the observations were made during the free-play period, which was judged to be the best time for observing spontaneous social interaction, the time for observing the different children was rotated from day to day according to a predetermined plan, so that no child would have an unduly large number of observations made during the "warming-up" period at the beginning of the hour or after possible fatigue had set in toward the close of the hour. In order to prevent similar bias from temporary changes in physical condition or from emotional disturbances, only one observation of each child was taken on a single day. Two general forms of behavior designated as *social participation* and *leadership* were studied in this way.

The procedure used was relatively simple. On the basis of preliminary observations, a series of descriptive categories for each of the two general forms of behavior was drawn up. At the beginning of the period, the observer, armed with stop watch and record blank, took up an inconspicuous position from which an unobstructed view of the subject to be observed could be had, and watched his reactions during a period of one minute.³ At the end of that time a decision was made as to which of the predefined behavioral categories best described the child's behavior during that particular moment and the fact was recorded on the blank. The observer then moved on to the next subject. The procedure was repeated

³ The preferred length of the observation period used in different studies by the time-sampling method has varied from five seconds to five minutes. For the most efficient use of time, the preferred duration is the shortest time required for adequate classification of the behavior. This will vary with the complexity and objectivity of the topic under investigation, and with the frequency of behavior changes. Time samples of physical activity, for example, must be very brief because the amount of activity displayed changes so rapidly from moment to moment. More time is required to decide whether a child is directing the activities of others or merely participating in their play.

day after day until the self-correlations for each of the separate categories were sufficiently high to meet the requirements of the problem.⁴

The categories used by Parten for the classification of social participation were as follows: (1) solitary with no observable occupation; (2) solitary play; (3) onlooker; (4) parallel group activity where each child carries on his own independent play but is a member of a physical group; a form of social behavior that is particularly characteristic of the two-year-old; (5) associative group play in which the children play together but without the assignment of individualized roles or group effort to attain a common goal; and (6) cooperative group play where all work together for a common end though not necessarily without friction, or each assumes an individual part in carrying out some dramatic activity.

Parten made no attempt to develop weights for these categories by statistical procedures as others have done since that time. Her major interest lay in studying the developmental trends for each category separately. She did, however, assign arbitrary weights to each category which enabled her to work out a roughly quantitative score for each subject. These scores correlated with the ratings of single teachers on a scale of social participation developed by Parten to the extent of about $+ .70$. For the pooled judgments of five teachers the correlation was $+ .88$. Since the forms of behavior here cited show a marked age relationship, and age was not held constant in determining the correlations, the figures cited are less significant than they appear.

The method of episode sampling is illustrated in a study by Murphy (1937), who observed the occurrences of sympathetic behavior as evidenced by indications of concern for another child in distress that were shown by two groups of nursery school children observed for 188 and 234 hours, respectively. Very detailed accounts of each episode were secured. Again the concern was chiefly with an analysis of the factors that give rise to sympathy and the manner and conditions under which it is displayed, rather than with an attempt to derive a measure of sympathetic behavior for the individual members of the group. Though more refined methods could obviously be devised, a simple counting method is all that has commonly been secured in studies of this kind.

The role of the individual in the group was also studied by Murphy. Every contact of each child with every other member of the group was noted, together with an indication of its general character. The results were summarized in graphic form as shown in Figure 32.

⁴ It has been found that the Spearman-Brown prophecy formula can be used to predict the approximate number of observational samples necessary to reach any specified level of self-correlation after a relatively small number of observations have been made.

In one of the pioneer studies of this kind, Olson (1929) used a method which is in some ways a combination of the two just mentioned. He studied the "nervous habits" of school children in the classroom, such as putting the finger into the mouth (including nail biting), fingering

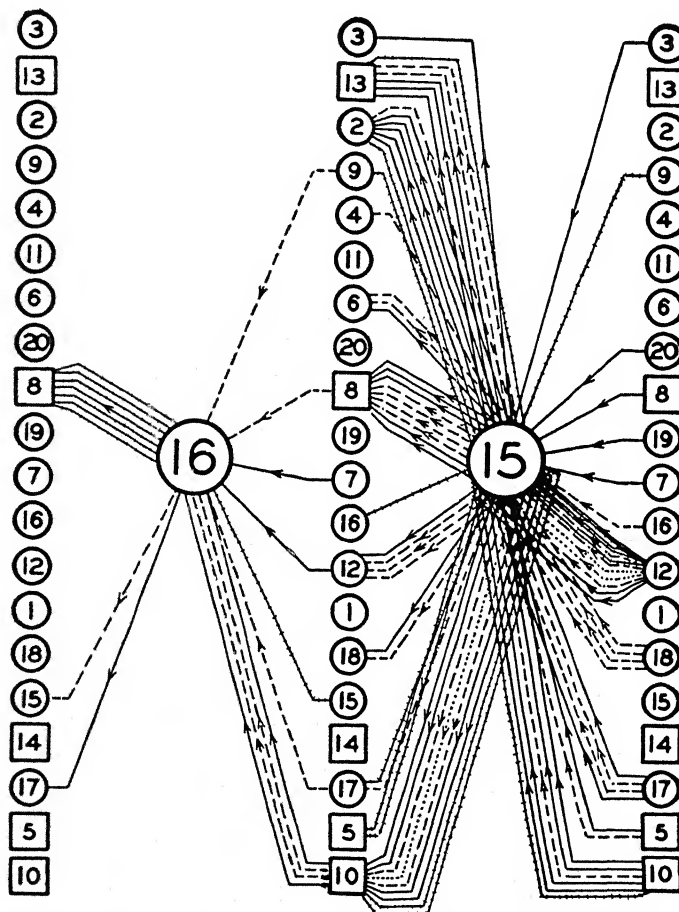


FIG. 32. DIFFERENCES IN THE NUMBER OF SYMPATHETIC SOCIAL CONTACTS MADE AND RECEIVED BY TWO NURSERY SCHOOL CHILDREN: DIAGRAM OF INDIVIDUAL ROLES IN THE GROUP. (Reproduced from Lois B. Murphy, *Social Behavior and Child Personality*, by permission of the author and of the Columbia University Press.)

the hair, or pulling the ear. An entire class was observed at one time. He used five-minute time samples taken in immediate succession, a procedure that unquestionably brought about a spurious increase in the self-correlations obtained by the "split-half method," since a broken fingernail

or an irritating hangnail, an itching scalp, or other temporary discomforts would increase the number of manifestations of such "nervous habits" much beyond their usual frequency for some cases. Twenty observations were usually taken. No attempt was made to count the actual number of such occurrences. Each child's score was taken as the number of periods during which one or more episodes of the kind under consideration was noted. A child who kept his finger in his mouth throughout a total period of five minutes received no higher score for that period than another who was seen to insert his finger into his mouth but once and that for only a brief instant. Other forms of classroom behavior, such as whispering or moving about the room, have been studied in the same way.

Inasmuch as speech is one of the main tools of social interaction, records of oral conversations between children or adults hold definite possibilities for the study of social traits. Most students of language behavior have been interested mainly in its symbolic aspects. They have studied the relative size of vocabulary or the average length of sentence, both of which have been found to be rather closely correlated with intelligence, as measured by standard tests. They have measured developmental changes in sentence structure and in articulation. They have studied sex differences, social class differences, and differences associated with the ordinal position of the child among his siblings. They have studied group differences in various linguistic functions such as the number and percentages of questions asked by the child, or the proportion of his remarks that have a definite emotional tone. But up to now, practically no attempt has been made to classify samples of a child's conversation taken under a variety of conditions in such a way as to provide an index, rough though it may be, of the characteristic features of his verbal contacts with others. The relative proportions of pronouns of the first, second, and third persons, remarks about people as compared to those about material objects, remarks about social institutions or abstract ideas, the frequency of complaints, criticisms, favorable or unfavorable comments, are among the reasonably objective aspects of conversation which, if properly analyzed, might throw considerable light on the social relationships of the individual. Volubility as indicated by the number of remarks in an hour of observation, as well as rate of speech, is also worth considering. That there are great individual differences in both of these factors has been demonstrated in a number of studies, but the significance of these differences for the individual has not been conclusively shown. The modern wire recorder has simplified the problem of securing the needed amount of data, a stumbling block in the way of making studies of this kind in the past.

Because the time-sampling procedure was developed at a time when a fairly widespread revolt against the so-called "artificial" conditions of formal tests and experiments with children as subjects was under way, the fact that the data were secured under conditions in which the behavior of the children could be regarded as natural and spontaneous won immediate popularity for the method as a means of studying social relations in childhood. It is less well adapted for use with adults, since it necessitates taking records at the time of observation without the knowledge of the subjects. Older children, as well as adults, are likely to become suspicious or at least unduly curious at the sight of notebook and stop watch, the more so if the observer makes an unsuccessful attempt at concealing them. Even with young children, a good many unanticipated difficulties and limitations of the method have been noted. In the first place, circumstantial factors are much more varied in the uncontrolled situation of the playground than under the more restricted conditions of the laboratory. This fact makes for greater freedom of behavior but adds to the difficulty of ascertaining what factors gave rise to it. The frequency with which different forms of behavior take place varies so markedly in different surroundings that nothing corresponding to the normative standards set up for tests and measurements of the usual kind can be established. It is possible to say with considerable assurance that during the free-play hour in the ——— Nursery School in February and March of 1948, Peter H. ranked lowest in social participation of all the children in the group. But in other surroundings, not only Peter's behavior but that of the other children might have differed enough to change his relative position very greatly. For example, in the study previously mentioned, Murphy found that in the two nursery schools she observed, the ratios of the number of instances of "sympathetic" behavior to the number of episodes classed as "unsympathetic" (such as ridiculing or teasing another child for crying when hurt), were respectively 1.63 and 6.63. She is of the opinion that the marked difference in the behavior of the two groups is to be attributed, not so much to any single factor or factors but rather to group interaction, to a kind of *esprit de corps* which is a function of group organization rather than of the individual characteristics of its members considered separately. The idea is plausible but would be more convincing if the descriptive anecdotes, of which many are given, had been supplemented by more adequate statistical analysis.

By securing a sufficient number of samples, time-sampling records can be brought to almost any required degree of self-correlation, provided that the situation under which they are obtained remains fundamentally unchanged. Their validity is in one sense intrinsic, but their

significance depends upon the extent to which similar results are obtained when the situation is changed. If a child is social in one situation, nonsocial in another, and actively antisocial in a third, each characterization being made in comparison to the typical behavior of others of his age and sex when placed in like situations, the possibilities of generalization from a single situation are limited. An adequate criterion of the functional validity of data obtained through direct observation of behavior outside the laboratory thus consists in a comparison of the variability (in standard score units) of individual behavior from one situation to another. If each subject maintains approximately the same relative position in the group, a change in the mean scores made by the group as a whole under different external settings does not affect the usefulness of a single series of observations as a means of classifying individuals with respect to their most typical behavior in the area under consideration (social participation, aggressiveness, sympathy, and the like). But if a change in the situation does not affect all subjects in much the same way, generalizations as to the behavioral tendencies of individuals on the basis of observations made under a limited series of conditions are likely to be seriously biased.

The great advantage of the time-sampling method for the study of social behavior lies in the fact that it makes use of real social behavior which is not easy to arouse in the laboratory, and provides a way of reducing the observations made to quantitative terms. Its limitations are: first, the difficulty of making records of the behavior of subjects who are no longer naïve in such matters; second, the large amount of time usually necessary for the securing of a relatively small amount of information; and, finally, the lack of dependable information with respect to the extent of generalization warranted by the kind of data usually obtained when this method is used.

METHODS OF STUDYING SOCIAL ORGANIZATION

In 1934, Moreno described a method of studying social organization to which he gave the name of *sociometry*.⁵ The procedures used are for the most part simple, though rather elaborate concepts have been built upon the results.

The fundamental principle in all sociometric theory is that no form of social organization can endure and no social group can remain stable unless the internal structure of the group is satisfying to its members. When a number of persons are thrown into temporary contact by exter-

⁵ Most of the studies in which this method has been used have appeared in *Sociometry*, a journal edited by Moreno.

nal forces, it is highly unlikely that all will unite to form a single organization, though certain individuals may seek each other out and so establish smaller groupings whose members find each other congenial. The internal organization of these groups may take on any one of many forms. Some will be organized closely about a single person whom all admire and whose friendship is coveted by all. Some take the form of a chain, where A seeks the companionship of B, who in turn is attracted by C. There are triangular or rectangular relationships where A seeks B, B seeks C, who in turn seeks A. There are mutual pairs in which each member prefers the other. These pairs may be isolated from the other members of the group, or one or both of them may also be sought as friends by other persons. Finally there are likely to be some who are either ignored or avoided by the rest.

In any free society, both the establishment of groups and their internal structure are determined chiefly by the wishes of their members. Since this is true, Moreno reasoned that a valuable approach to the study of larger groups might be had by an examination of the forms of social organization voluntarily set up by children and adolescents when free choice is allowed.

Figure 33 illustrates a very simple experiment of this kind. In this case, each member of a group of twenty boys, all of whom were cottage mates in a small private school, was asked to choose whom he would prefer to have as roommate. They were assured that their choices would be carried out as far as possible. This is an essential feature of sociometric technique, inasmuch as the attitude toward real alternatives is likely to be quite different from a mere pencil-and-paper statement which has no further consequence. In the actual experiment, each boy indicated a second and third choice. In Figure 33 only the first choices have been shown in order to simplify the diagram.

In this figure, known as a *sociogram*, the position of each boy in the general group structure is indicated by an identifying letter. The direction of the arrows shows each boy's choice. Thus, A chose O but O preferred N. B preferred C, whose choice was P, and so on.

Moreno noted five main types of social organization, all of which can be seen in Figure 33.

The *isolates*. These are cases not chosen by anyone else. In Figure 33 they are subjects A, B, H, K, Q, and S. Their number would undoubtedly have been reduced if second and third choices had been indicated on the sociogram.

Mutual choices. Cases E and D; J and N; M and L. It will be noted that the first pair differ from the other two in that although each boy

makes the other his first choice, neither is preferred by any other member of the group. This obviously calls for further study.

Triangles or rectangles, such as the relationship extending from I to G, from G to R, and from R back to I.

Chains, such as the line from Q to T to F.

Stars. These are the very popular subjects whom many choose. Cases J and F are the best examples here shown.

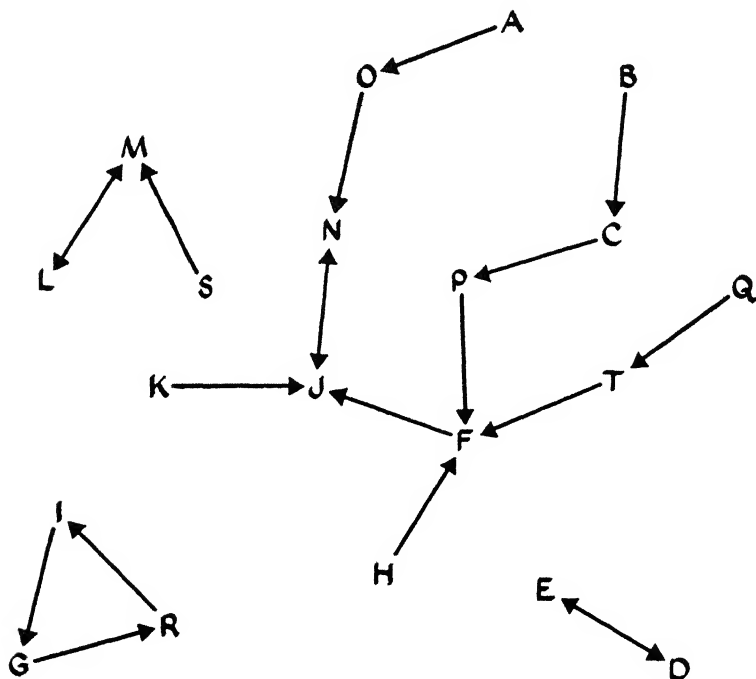


FIG. 33. SOCIOGRAM SHOWING CHOICES OF ROOMMATE BY EACH OF TWENTY BOYS IN A SMALL PRIVATE SCHOOL.

It is evident that the sociometric procedures hold definite possibilities both for the scientific study of social structures and for the identification of individuals whose patterns of social relationship are such as to merit special study. What factors, for example, differentiate the isolates from the "stars"? Does Q choose T as a possible means of reaching the very popular F? What factors make for the apparent isolation of the two small groups formed respectively by D and E and by G, I, and R? Many other similar questions are suggested.

It is obvious that more information than is shown here is needed for a valid sample of the social relationships among a group such as this.

In addition to the second and third choices of roommate, preferences for companions in other situations or activities must be considered. Both preferences and rejections arise from many causes. The most popular person for one activity may be rejected when the choice has to do with another and quite different situation. A study of the factors which make for such differences in the social role of a given person may often throw considerable light on the nature of his social relations and on his personality characteristics as well.

The sociometric position of a given person is likely to change as the result of social experiments in which he plays a part. The roommate chosen today may be rejected after a few months' trial. These changes, too, are significant for the study of the individual persons concerned as well as for sociological theory.

OTHER METHODS OF STUDYING SOCIAL RELATIONSHIPS

As commonly used, the experimental method is not well suited to the study of individuals, both because of the difficulty of stimulating genuine social behavior in the laboratory and because of the inconvenience of securing a sufficient amount of data to be dependable if the sampling method is relied upon. Inasmuch as two or more individuals are involved in each situation, and every change in the composition of the group means an alteration in the situation for the subject under consideration, the amount of data needed for an adequate sample of his social reactions is far greater than is required for a test of ability where the situation can be kept constant from trial to trial and the subject's behavior can be directed by straightforward instructions as to what he is to do. An unpublished study by Marjorie Walker⁶ provides some evidence on this point. Each member of a group of fourteen nursery school children was paired with every other member in three different situations, making a total of 273 trials for each child. The behavior on each occasion was recorded by an observer stationed behind a one-way vision screen. It was found that in spite of the unusually large amount of data secured, only rather moderate self-correlations for the individual records were found. The behavior of Child A when paired with B was likely to vary considerably from that shown when he was paired with C. Chance factors which vary with the particular occasion exert their maximal effect when no instructions are given.

Some attempts have been made to utilize signs instead of samples

⁶ Ph.D. thesis on file in the University of Minnesota Library.

as a means of securing evidence as to the social tendencies of an individual. Moore and Gilliland (1921) and Gilliland (1926) devised a method of measuring "aggressiveness" by means of an experimental situation in which the effect of staring the examiner in the eye while performing set tasks of mental addition was one of the measures used. Scores were based upon a comparison of additional time required and increase in the number of eye-movements in the experimental situation as compared with a control series. A test of changes in rate of writing under pressure of time and a test of word association were also included. The authors report self-correlations in the neighborhood of $+ .60$, and correlations with ratings by associates of about the same level for small groups of college students. The test has not been much used except by its authors.

A measure of body sway has been used as an indication of the readiness of response to suggestion and particularly as an indicator of susceptibility to hypnosis. The procedure described by Hull (1933) gives a very accurate determination with self-correlation for unselected college students well over $+ .90$. Correlations with ratings on suggestibility as well as with other more elaborate tests designed for the same purpose have been uniformly positive and moderately high. There is evidence that it is also a reasonably good indicator of neurotic tendencies.

A number of physiological measures have been employed for the study of the bodily changes associated with emotional stress. One of the best known of these is the "lie detector," which makes use chiefly of changes in systolic blood pressure. Changes in the electrical resistance of the skin as measured by the string galvanometer, or in the distribution of blood as shown by various types of plythsmograph, measures of muscular tension or muscular tremors, and various other physiological changes have been employed, not only for the identification of temporary emotional states but also in the attempt to develop methods that might prove useful for individual diagnosis of more basic and lasting personality characteristics. Up to the present time, these attempts have not been highly successful.

Decidedly more fruitful has been the search for "signs" of personal-social characteristics in measures that overtly deal with other matters, such as interest in fields that apparently have no relationship to the characteristic in question, or the particular pattern of performance on tests of intelligence. Among the former are the interest-attitude scales devised by Terman (1938) and by Burgess and Cottrell (1939) for the prediction of happiness in marriage, and the Terman and Miles (1936) scale for measuring mental masculinity-femininity characteristics. Furfey's (1931) scale for measuring what he calls "developmental age" is

another example of this kind. All these measures have shown high self-correlations and some relationship to the characteristics which they were intended to measure. The difficulty of securing suitable criteria outside the tests themselves, particularly in the case of the "masculinity-femininity tests," has been an unsolved problem except for such extreme deviates as manifest homosexuals whose test scores have usually shown very marked conformity to actual behavior.

Among the attempts to find signs of personal-social characteristics within the pattern of performance on a test designed for another purpose have been those using the Wechsler-Bellevue Scales for measuring intelligence. The five signs or clinical syndromes so far worked out are described by the author in *The measurement of adult intelligence* (1944). Each of these syndromes is said to be suggestive of one of the following conditions: (1) neurosis, (2) organic brain damage, (3) schizophrenia, (4) psychopathic personality, or (5) mental deficiency. Evidence for the validity of these claims is conflicting.

Attempts have also been made to utilize a sampling method, similar to that found so successful in the measurement of intelligence and educational accomplishment, for the prediction of personal-social conduct. By far the most elaborate of these is the series of studies on the measurement of deceit, and the prediction of such deceitful acts as stealing, cheating, and lying, which were carried out by Hartshorne and May (1928). The same authors together with J. B. Maller also devised a long series of tests designed to measure various aspects of service and self-control (1929). These studies have been so frequently described that only brief comment is called for here. The net result was to show (1) that the sampling method in which tests of actual conduct of the kind to be investigated are utilized is capable of yielding information of considerable value for individual diagnosis but (2) that the amount of data required for an adequate sample, the necessity for using not only a variety of tests but also a variety of settings in which the tests are to be employed, as well as the cumbersomeness of the procedures and the excessive amount of time required for them render the method of sampling by means of concrete situations not feasible for ordinary use. A further objection lies in the fact that many of the tests used, such as those which provide not only opportunity for cheating but also strong motivation to cheat, are open to definite criticism on moral and ethical grounds.

Of all the performance tests of this kind that have been tried, perhaps the most promising is the overstatement test, which has been tried out by a number of persons and most thoroughly by Raubenheimer (1925). As a paper-and-pencil test, this is easily administered in the

ordinary classroom without special apparatus or unusual conditions. It is largely self-motivating since it appeals to the natural desire to make a good showing. Time for each item is no greater than that required by those used in the ordinary group intelligence test; hence a sample of sufficient size and variety can be secured within a reasonable period. Inasmuch as the test measures both understatement and overstatement, the scores earned cover the entire range from excessive lack of self-confidence to unwarranted boastfulness. The device is one that merits more study than it has received.

An overstatement test consists of two parts. In the first part the subject is asked to say whether or not he knows or can do certain things. The second part, given immediately after the first, is a test of his actual ability along those lines. Each part is scored separately and the difference between the two scores is then converted into a percentage of the amount of overstatement or understatement made possible by the extent of the subject's knowledge of the facts in question and the limits set by the test.⁷ The first part would include such items as the following:

Do you know who was the first president of the United States? . . . Yes—No
 Do you know what a cypress is? Yes—No
 Do you know the multiplication table up to 12×12 ? Yes—No

The second part would include items similar to these:

The first president of the United States was:—Christopher Columbus; George Washington; Abraham Lincoln; Paul Revere.

A cypress is a kind of:—tree; cloth; dog; food.

$8 \times 9 =$:—48; 96; 56; 72.

If the score indicated by the subject's claims is larger than that determined by the test of his actual knowledge, the difference is given a plus sign indicating overstatement. Inasmuch as the greatest amount of overstatement possible under the conditions of the test is the difference between the child's score on Part II and a perfect score on that part, the number of points by which he overstates his knowledge is divided by the difference between his score and the maximum score. The result is called his *per cent of overstatement*. If, on the other hand, the score on Part I is smaller than that on Part II, the *per cent of understatement* is found by dividing the difference between the scores on the two parts by the score on Part II and prefixing a minus sign. By reducing the figures to percentages, allowance is made for differences in the amount of information possessed by the different subjects for whom the possible amount of

⁷ Overstatement is possible only with respect to that which one does not know; understatement only with respect to what one does know.

overstatement or understatement permitted by the test would otherwise be unequal.

By far the most elaborate attempts to utilize signs for the study of personal-social characteristics are to be found in the so-called "projective" methods. Although these procedures have been chiefly employed for the study of the affective and attitudinal side of mental life which we have here designated as "personality," the line between social and personal relationships is at best very tenuous. The theory underlying the projective technique is, however, so different from the more conventional methods of approach that we shall reserve its discussion for the next chapter.

STANDARDIZED RATING SCALES FOR APPRAISING SOCIAL BEHAVIOR

The Vineland Social Maturity Scale prepared by Doll (1936) is, strictly speaking, neither a rating scale nor a test, but falls midway between the two. It consists of a series of items arranged in the form of a year scale. The subjects are not tested directly. The information is supplied by parents or others in a position to know the child's abilities. The items include information regarding (1) self-help, (2) locomotion, (3) communication, (4) occupation, (5) self-direction, and (6) socialization. The scale is particularly useful in work with mental defectives inasmuch as it provides information on certain practical aspects of mental development which sometimes show marked divergence from those measured by the usual tests of intellectual capacity. It has also been found to be a much-needed help in parent education since it provides an objective standard to which reference can be made in the very frequent cases in which parents of normal children are either attempting to hold them to standards that are unreasonably high for their level of development or are failing to help them to acquire the skills of which they are capable.

A series of rating scales prepared by the staff of the Fels Institute for Child Research has been described by Champney (1941). Ratings are made by the home visitor after training in the method. Usually several ratings by different judges are secured. Thirty separate aspects of parent-child relationships are rated, including such factors as "sociability in family," "amount of disciplinary friction," "severity of punishments," and the like. Six levels, ranging from a high degree of the behavior in question to little evidence or only rare manifestations of it, are distinguished. Each level is carefully described.

A number of other rating scales following the usual graphic form

have been devised for rating personal-social behavior. As they do not involve any extraordinary features, they will not be described here. Mention should perhaps be made of the Read revision (1940) of the Conrad Behavior Inventory for Nursery School Children (1933) of which the distinctive feature consisted in utilizing persons with much experience in the practical management and training of young children to help in the selection of items deemed most significant for the description of child personality. In building a rating scale the psychologist is likely to be guided largely by theoretical and statistical considerations. Particularly when, as in the case under consideration, no really objective criterion is available for judging the importance of a given behavior tendency in the child's personality make-up, in his attempt to devise a measure that will really represent the area of behavior in which he is interested, he is likely to include a good many items that are of little practical consequence. Conrad's original series of ratings included 231 scales. Its length tended to preclude its use in many institutions and often led to hasty and ill-considered judgments on the part of busy teachers who were required to supply the information called for. By means of a careful screening procedure, Read was able to reduce the length of the inventory from 231 scales to 67 without material loss.

PAPER-AND-PENCIL TESTS AND QUESTIONNAIRES FOR THE APPRAISAL OF PERSONAL-SOCIAL CHARACTERISTICS

Whether or not the average person can or will give a truthful account of his own feelings and experiences with respect to intimate matters which he is accustomed to look upon as highly personal and private affairs, or whether, regardless of their factual accuracy, the statements he is willing to make about these topics have sufficient validity as signs from which useful information about his personal-social characteristics may be drawn, is a moot question. The multiplication of "personality inventories," "temperament tests," and the like which are based upon self-report, as well as the many experimental articles dealing with the results obtained by their use, suggests either that many people continue to find these tests useful or that they have been very effectively deluded into accepting the results as valid. The statistical evidence to warrant the uncritical interpretation at their face value of scores obtained from these devices, however, is not always convincing, although the recent scathing condemnation of the entire group of questionnaires of this kind by Ellis (1946), whose general attitude appears to be "a plague on *all* your inventories," is perhaps unduly severe.

The general method of the personality inventory does not differ greatly from one published series to another. The selection and wording of the questions vary, though there is much overlapping, and the special aspects of personality of which samples or signs are presumably afforded by the different inventories differ in accordance with the special interests of the authors. In some of the inventories the questions⁸ are to be answered in terms of an unqualified "yes" or "no"; in others a graded series of answers, such as "always," "often," "occasionally," or "never," is provided for. Sometimes statements are substituted for questions in the belief that a somewhat more remote formulation may be less likely to arouse antagonism. In such cases the subject is asked to say whether or not he agrees with the statement. Again, either a "yes" or "no" answer or a graded series of replies is provided for checking. A number of other modifications of the direct question have been advocated, but little dependable evidence has been adduced to show the relative advantage of one procedure over another.

An exception to the statement just made is to be found in a report by Rundquist and Sletto (1936), who made a careful study of the effect of the form of statement upon the significance of the replies to it. It was found that when a statement concerning a given attitude or belief was so worded that agreement with it would conform to the pattern generally regarded as socially desirable, many people whose actual opinion was not in conformity with the statement would nevertheless not bother to disagree with it. Disagreement, they found, is in general much more significant than agreement, especially when it runs contrary to conventional ideas of social acceptability. It follows that when the items in a personality inventory are worded in the form of statements, the responses will be more revealing if these statements provide a real challenge to thought. For example, many persons who, in the course of responding to a long list of statements concerning their social and political beliefs, express a hasty agreement with the conventionally worded statement, "I believe that democracy is the best form of government," will be brought up short by the statement, "I do not believe in democracy as the best form of government." The difference seems to lie in the type of response called for when a given opinion is expressed in both responses. The affirmative response to the first question calls for nothing more than a passive acceptance of an attitude

⁸ Examples of the kind of questions usually asked are the following:

"Are you happy most of the time?"

"Do you sometimes feel like jumping off when you are on a high place?"

"Do you usually feel well and strong?"

"Do other people often treat you unfairly?"

widely regarded as "correct"; it may or may not indicate a personal conviction based upon careful consideration of the evidence. But the second question immediately offers a challenge to conventional attitudes, at least in the United States. While it is still possible that many who disagree with it are still responding on a rather superficial level, they are at least forced to take an active rather than a merely passive role in so doing. This is probably the basis for the finding of Rundquist and Sletto that responses which run contrary to the socially accepted form of a question are more likely to be valid indicators of well-established beliefs and attitudes than are those of a more conventional type.

A point that is not answered by the Rundquist and Sletto study is this: If *all* the items in such a questionnaire are phrased in terms of negative acceptability, rather than just a few as was true in their study, is it likely that an accumulative attitude of opposition would be built up which would negate any original advantage of the form of wording employed? Until this question has been answered, it is unsafe to advocate general adoption of the device.

That the personality inventory, when carefully constructed, is capable of yielding results that are internally consistent has been repeatedly shown. Self-correlations, either by the split-half method or as determined by retests after a short interval of time, usually fall short of those commonly found for tests of ability, but are nevertheless high enough to meet any reasonable test of statistical significance. The meaning of this, however, is far from being self-evident. Some of the critics of the method have rudely suggested that this consistency of self-report merely affords one more demonstration of the fact that the person who tells a lie is likely to try to stick to it. Others have suggested that the only characteristic really measured by these inventories is a peculiar kind of exhibitionism which takes the form of an urge to display one's troubles and confess one's weaknesses. Another possibility along the same line is that exceptionally high scores⁹ may signify an overdeveloped conscientiousness, a meticulousness with respect to small matters that approaches if it does not completely reach a state of mental abnormality. It is quite possible that one of the reasons for the apparent superiority of the Minnesota Multiphasic Inventory, constructed by Hathaway and McKinley (1940),¹⁰ over many of the other questionnaires of this kind is

⁹ Personality inventories are usually scored in terms of the number of undesirable responses made. High scores are accordingly "bad"; low scores, "good."

¹⁰ Accounts of the different scales included in this inventory have appeared in a number of psychological and medical journals. References to these reports and to additional studies in which the inventory was used are given in the very complete bibliography by Ellis (1946). Materials and instructions for its use can be secured from The Psychological Corporation, New York City.

the fact that the method is better adapted to the construction of a scale for the identification and classification of cases of mental disorder in its early as well as in its later stages than to diagnosing the personality characteristics of normal people. It may well be that the maintenance of a decent reserve with respect to the more intimate affairs of one's inner life is the normal way of behaving, and that the person who is willing to reveal all at the first opportunity exposes his mental aberration by his very readiness to bare his troubles to the world.

That the responses to these questions should be regarded only as "claims" and not as objective facts is coming to be generally recognized. As claims they may be useful signs, regardless of their truthfulness. It then becomes the task of the experimenter to ascertain the direction in which the sign points. He need not be particularly concerned with finding out whether or not A—who gives an affirmative response to the question, "Do you usually sleep well?"—actually does sleep longer hours and more quietly than B, whose response is negative. He is concerned only with what is claimed and with what may be inferred from that claim.

The Multiphasic Inventory consists of 550 items. The authors recommend its use as an individual test, but forms for administration as a group test are also available. In the individual form, each question is printed on a separate card. The subject sorts the cards into two piles, accordingly as his answers to the questions are affirmative or negative. The great advantage of this procedure lies in the fact that the subject does not realize that any permanent record of his replies is made.¹¹ He is thus more ready to respond in a relatively free and uninhibited manner, since with the possible exception of the examiner no one, so he thinks, will know what he has told.

The test was devised and standardized in a psychiatric clinic attached to a mental hospital. Eight scales have been derived from it, each of which is calibrated in standard-score units (T-scores). These are (1) hypochondriasis, (2) depression, (3) hysteria, (4) psychopathic deviate, (5) masculinity-femininity, (6) psychasthenia, (7) schizophrenia, (8) hypomania. An additional device called the K scale has been worked out by means of which a correction is made to the raw scores when there is evidence that the claims made are either excessive or insufficient because of the subject's attitude toward the test situation. Meehl and Hathaway (1946) and McKinley, Hathaway, and Meehl (1948) have given the rationale and described the procedure for making this correction. Their figures indicate that its use tends to raise the

¹¹ This statement may not hold good for the more sophisticated subjects, but it is true of the majority.

intercorrelations between scales, as would be expected if K represents a generalized attitude toward the situation as a whole. Not all the scales are equally affected by the K factor, however, and certain ones show so little effect that its use in those cases is not recommended. The authors believe that some improvement in the diagnostic usefulness of the five scales in which the K corrective is applied is thereby accomplished.

The Multiphasic is probably the most carefully standardized of the personality inventories that has been devised up to the present time. Among the others are the Bernreuter (1931), which has probably been more widely used than any other. It is designed to measure four different aspects of personality: neuroticism (B₁-N), self-sufficiency (B₂-S), introversion-extroversion (B₃-I), and dominance-submission (B₄-D). Although Flanagan (1935), by the use of a factorial analysis, showed that the intercorrelations among Bernreuter's four traits can be largely accounted for by two underlying factors designated by Flanagan as self-confidence (F₁-C) and sociability (F₂-S), most people have continued to make use of Bernreuter's original classification. The Humm-Wadsworth Temperament Scale (1935, revised 1940), the Bell Adjustment Inventory (1934), and the Bell School Inventory (1939) have also been used extensively. Many others of the same general type are on the market.

The use of factorial methods has done much to clarify thinking in the field of personality measurement, even though tests designed to measure the factors that have been isolated in this way have not proved universally superior to others that make use of self-report. Guilford (1939), for example, was able to show that the much-discussed "trait" known as introversion-extroversion is not a unitary characteristic but has at least five components designated as (1) social behavior displayed in the tendency to seek companionship rather than solitude, (2) type of thinking, as shown in preoccupation with external events and activities as opposed to abstract ideas, (3) depression versus elation, (4) cycloid tendencies as opposed to stability of mood, and (5) rathymia, as displayed in the contrast between a typically carefree, "happy-go-lucky" type of disposition and a serious-minded, overly conscientious outlook upon life. That the various scales designed to measure introversion-extroversion have typically yielded very different results for the same subjects is in all probability due to the fact that some emphasize one of these aspects while others place more stress upon one or more of the others. The use of a common name by no means guarantees similarity of meaning. Other concepts for which the inventory methods have been employed, such as dominance-submission, vocational adjustment, and

the like, have been analyzed in a similar manner and have likewise been found to be based upon multiple rather than unitary factors.

The monograph by Raymond Cattell (1946) is undoubtedly the most thoroughgoing attempt that has yet been made to utilize factor analysis as a means of charting the entire structure of the personality. Cattell points out that personality can be defined only in terms of the fields with which it is concerned. He states (page 566) that "personality is concerned with and deduced from all the behavior relations between the organism and its environment. It is that which predicts behavior, given the situation."

The attributes by which it is described and measured are traits (structures or dispositions defining potential behavior) which may be considered properties of the organism but which can only be defined in terms both of the organism and of its environment, i.e., "as relationships between the physical organism and its environment."

Cattell began his analysis with an examination of the 17,953 trait names listed by Allport and Odbert (1936). After disposing of synonyms or near synonyms and adding a few from other sources, he came out with a list of 171 that did not appear to duplicate each other as far as surface inspection permitted him to judge. A search of the literature revealed some fourteen studies in which correlations had been worked out between ratings on various combinations of these traits for different groups of subjects. The most nearly complete of these was one that Cattell himself carried out in connection with the study.

By examination of these correlations he was able to find certain groups of traits, designated as *clusters*, in which each one correlated with all the others to a significantly positive degree. From these he selected smaller groups to which he gave the name of "nuclear clusters" because they appeared together in a number of different larger clusters.

A group of 208 men was divided into smaller groups of 16 each in which all the members were well acquainted with each other. Ratings of each man by all the other members of his group were secured on each of the basic traits entering into his nuclear clusters and a factorial analysis was made of the results. From this there emerged a list of twelve factors which were tentatively named as follows: (1) cyclothymia vs. schizothymia; (2) intelligence vs. mental defect; (3) emotionally mature stable character vs. demoralized general emotionality; (4) dominance vs. submissiveness; (5) surgency vs. agitated melancholic desurgency; (6) sensitive, anxious emotionality vs. rigid, tough poise; (7) trained, cultured, socialized mind vs. boorishness; (8) positive character integration vs. immature, dependent character; (9) charitable, adventurous cyclothymia vs. obstructive, withdrawn schizothymia; (10) neurasthenia

vs. vigorous "obsessional determined" character; (11) hypersensitive, infantile sthenic emotionality vs. phlegmatic frustration tolerance; (12) surgent cyclothymia vs. paranoia.

Each of these traits is described in detail in Cattell's monograph. When they seem to correspond fairly well to the characteristics presumably measured by some of the well-known tests, that fact is noted. In each case, examples of descriptive ratings that were found to be heavily loaded with the trait under consideration are listed.

Not everyone will agree with Cattell's point of view. Some will object to his rather stilted terminology, in spite of the fact that his descriptions of the twelve traits just named are for the most part clear and sufficiently vivid.¹² But certain points in his discussion are of such fundamental importance for the entire field of personality measurement that they can hardly be too often repeated.

First, there is the need for defining and charting the field to be studied. For more than half a century psychological investigations on the personality characteristics of man have been carried on with little attempt at systematic exploration. Much, it is true, has been learned in this way. Instruments have been devised that have considerable usefulness for certain purposes. But on the whole, the method of search has been much like that of a flock of chickens when one of them finds a worm. Immediately the entire flock rushes frantically to the spot, all hoping to find more worms in the same place. What is needed is a more systematic covering of the field, based upon accumulating knowledge of its most fertile areas. Cattell's outline may prove to be as far removed from the actual facts as were the maps of the world prepared by the early cartographers, but in any case it is an outline, based upon such data as are available and organized according to a procedure that is at least straightforward. One may disagree with the underlying theory. One may disapprove of the method used. Something is nevertheless gained by the provision of an internally consistent plan, the errors in which may be checked. Cattell makes no claim that his results are infallible or that other approaches to the problem may not prove to be more fruitful than his own. But he does feel very strongly that the time for random experimentation has passed and that the work of the future should be guided by a more systematic concept of the goals to be sought.

A second important feature of Cattell's theoretical position is the emphasis he places upon the *situation* as a basic attribute of the *response*. He unqualifiedly rejects the idea that a trait can be defined *in esse*.

¹² For definitions of these terms, see the Glossary.

For him it must always be *in posse*, manifesting itself in terms of the conditions immediately present. The intelligent person does not manifest his superior ability at all times and on all occasions but only as circumstances call for it. The cyclothymic individual exhibits changes of mood in response to circumstances which would leave his more phlegmatic neighbor undisturbed. Although internal as well as external factors may give rise to these shifts, they can only be predicted in terms of the relationship between the person and the situation in which he is placed.

Allied to the above principle is Cattell's discussion of the interaction between traits. In the brief account of factor analysis given in a previous chapter of this book, it was noted that the interaction of factors may prove to be quite as important a field of investigation as the study of the factors themselves, inasmuch as the mode of operation of the separate factors may be so greatly changed in accordance with their organization with respect to each other. Intelligence is generally regarded as an asset, but an intelligent criminal is likely to be far more of a menace to society than is a stupid one. Submissiveness is likely to be a desirable trait in an individual who is physically or mentally incapable of caring for himself. Social ascendance is a valuable asset to the intelligent and emotionally stable person, but in the self-centered person with paranoid tendencies it may become a dangerous gift.

Also worthy of mention is the distinction which Cattell draws between what he calls "source traits" and "surface traits." The latter can be observed directly; the former are the more basic but can be discovered only by mathematical analysis and psychological insight. Thus the source traits correspond to factors; the surface traits are the forms of behavior to which these factors give rise. For example, *interest in problem solving* would be one of the cluster of surface traits associated with the source trait *intelligence*; *self-control*, with the source trait *emotionally mature, stable character*. The number of surface traits is obviously greatly in excess of the number of source traits. Because surface traits are open to general observation, it was but natural that they should be the first to stimulate attempts at measurement, but their number is so great as to preclude an effective appraisal of an individual in such terms.

Cattell has not as yet provided a series of tests for the measurement of his twelve source traits. This is a more difficult problem and one that will provide the final test of the soundness of his general concepts and the efficiency of his methods. The question here extends far beyond the particular formulation arrived at by Cattell. At least three more general problems are involved. First, there is the question as to whether the rating method is to be preferred to the indirect methods based upon signs,

such as are obtained from tests of interests and attitudes in which special scoring keys are used, or by means of the various "projective techniques" described in the following chapter. Second, there is the question whether the methods of factor analysis as used by Cattell, while they may be well suited to the ferreting out of general relationships that hold good on the average, are capable of adaptation to the widely variant conditions encountered in individual diagnosis. Finally, one may ask: Is the entire concept of "traits," as ordinarily regarded, in need of rather drastic overhauling?

Because the only one of the source traits in Cattell's list for which reasonably satisfactory measuring devices are at present available is intelligence, it may be well to look there for suggestions. That intelligence cannot be satisfactorily measured without trial is a truism. In order to ascertain the mental capacity of an individual, we set tasks of increasing difficulty and provide incentives whereby he is encouraged to try to master them. Intelligence, as thus conceived, is the *ability* to perform certain types of mental tasks when properly motivated to do so. The confusion between ability and conduct was for long a stumbling block in the field of intelligence testing, as it still is in the measurement of other personal-social characteristics. The emphasis which Cattell has placed upon the importance of the situation as an essential aspect of the definition and description of a *trait* does not seem entirely consonant with his use of the conventional type of ratings which take no account of circumstances as a basis for appraising the individual. Perhaps a more serious attempt might be made to handle the vexed problem of the measurement of personal-social traits as *abilities* rather than as tendencies to certain forms of behavior as Cattell and most others have regarded them. Thus we should have the *ability* to respond to changes in the external situation by appropriate changes in mood (cyclothymia), the *ability* to dominate, the *ability* to "come back" after painful or disturbing experiences (surgency), and so on. The assumption would then be made that each individual possesses all of these basic or "source" characteristics but in varying degree, and the problem would be to find appropriate tasks to constitute a series graded with respect to difficulty for the measurement of each. This is not an easy matter, but our long-established habits of thinking in terms of surface traits may be in part responsible for the difficulty. One thing appears certain. The practice of overlooking the factor of motivation in the study of personal-social behavior has been a vicious source of distortion in most of the studies up to the present time.

Projective Methods for the Study of Personality

FUNDAMENTAL CONCEPTS

That all of us tend to ascribe our own inner feelings and sensations to objects in the world outside our own bodies is well known. We say that our soup is cold, meaning that it does not arouse the desired sensation of heat in our mouths. We speak of an annoying person, meaning one who causes us annoyance. We blame others for our own mistakes.

Not only in speech but in behavior is the same tendency shown. Actions are frequently a form of language by which the individual unconsciously betrays his thoughts, his feelings, and his emotions. The office manager, usually a genial and tolerant person, may become curt and unreasonable during an attack of indigestion. He may find fault with his secretary, discharge the office boy for some trivial fault, and raise a major tempest over a misfiled letter. He *projects* his own feelings onto the persons and things surrounding him.

In some forms of mental disease, such projection is carried to the point of actual delusion. The external world becomes peopled with the creatures of the patient's own distorted fancy. He sees demons or hears the voices of angels or of long-dead friends. He is terrified by ghosts, by dangerous reptiles, by malevolent enemies. Sometimes these delusions have no apparent reference to objective reality. Often, particularly in the less advanced cases, they are referred to some person or object which is transformed into the shape dictated by the disorganized inner life of the person concerned.

The basic theory underlying the use of the so-called "projective methods" is that, because of differing innate tendencies which have become still further differentiated as a result of modification through diverse experiences, no two persons will perceive the world in exactly the same way. Because their perceptions differ, their responses to these perceptions will also differ. The behavior of every person thus provides

the observer with a series of signs which, if properly interpreted, will enable him to understand much of the thoughts and feelings, the hidden motives and un verbalized meanings which Frank (1939) has so cogently designated as the "private world" of the individual.

The *projective method* may be described as a system of diagnosis based upon signs rather than upon samples. The usual procedure consists in presenting the subject with some kind of standardized material such as toys, paints or other art material, pictures, inkblots, and the like. The choice of material will depend to some extent on the age of the subject and on the overt nature of his difficulties. However, most of the workers in this field seem to have developed definite preferences for one or another of the methods in common use, and employ that one to the practical exclusion of the others.

In general, the materials used and the instructions given provide what is known as "slightly structured" situations. That is, the materials are such as to suggest certain uses rather than others, while some additional specificity is provided by the examiner, who indicates in a general way what is to be done with them but is careful to avoid all suggestion as to the manner of its doing. Usually the subject is specifically told that there are no right or wrong answers, no set ways of dealing with the material. He is to handle it in whatever way he sees fit or in accordance with his own wishes. Children are simply told that they may do whatever they like.

Various ways of recording the behavior are used. When the result is a material product such as a drawing or painting, this product itself forms the main part of the record and is preserved for further study. While the work is in progress the examiner takes a few inconspicuous notes on the subject's behavior, giving particular attention to spontaneous comments and to evidence of bodily tension or other signs of emotional disturbance. At its conclusion it is customary to question the subject about any special or unusual features of the work, or any parts whose meaning is not clear. This practice of subsequent questioning is coming to be regarded as an important feature of the projective technique, regardless of the approach used. When only verbal behavior is involved, recording is less easy although the use of mechanical recording devices has simplified the problem when these are available. Sometimes a stenographer seated behind a screen takes down the record. If none of these aids can be had, the examiner is forced to take his own notes. Some depend upon memory for this, making the record directly after the interview is ended. Others try to take notes surreptitiously, a plan that is rarely successful except (perhaps) with small children. Generally speaking, if the note taking must be done by the experimenter and if

the notes are presumed to constitute an exact record of what went on, it is better to take them openly, making whatever explanation is necessary to satisfy the subject.

A complete account of the methods used would greatly exceed the limits of this chapter. Bell (1948) has reviewed the experimental literature on a number of the most popular devices very adequately and has added a well-selected bibliography on each method which the interested reader will do well to consult. Here we shall attempt nothing more than a brief description of a few of the more widely used procedures of this kind, together with a few notes on their possibilities and limitations. By way of introduction a brief account will be given of what is perhaps the earliest and certainly one of the best studies of this kind that has appeared in the literature—Binet's *L'Étude expérimentale de l'intelligence* which appeared in 1902.

BINET'S EXPERIMENTAL STUDY OF INTELLIGENCE

Throughout his life, Binet's concept of intelligence was broad rather than narrow. He was interested in the qualitative aspects of thought and behavior. "In what manner?" was for him quite as important a question as "How much?" Indeed, had it not been for his careful and painstaking observations of the differences in the *way* children of successive ages approach the problems presented for their solution it is questionable whether his intelligence tests could ever have taken the form that they did.

By the turn of the century, Binet's work had attracted enough attention to provoke criticism from some who were most strongly wedded to the system of psychology established by Fechner and Wundt. Binet's methods, they said, were not scientific; they lacked the precision of measurement made possible by the use of more elaborate apparatus. Moreover, it was claimed that the problems with which Binet dealt had to do with matters that are, by their very nature, unknowable, at least until the groundwork has been laid by a very thoroughgoing study of the more basic processes of mind, particularly the sensations and the stimuli that give rise to them.

In reply, Binet stated that the work then being carried out in the German laboratories was primarily physiological rather than psychological, for it had borrowed from physiology its apparatus, its methods, and its stimuli. Attention had been centered upon the material conditions of experiment rather than upon the subjects who participated in it, and who were, in effect, regarded as parts of the apparatus rather than the primary objects of study.

Binet pointed out that the term "stimulus" had been too narrowly understood, even by such men as Ribot. The immediate stimulus presented in the course of a laboratory experiment is only a very small part of the total. Every experience to which an individual is subjected sets up an elaborate and complicated series of reactions that persist and interact with those occurring later. It is this complex, not the limited part of it presented at a given moment, that is the real stimulus.

Binet emphasized the important role of language in this process. The fact that a verbal symbol may be as effective as a physical object is one of the reasons why elaborate apparatus is not always necessary for truly scientific work. Binet did not, by any means, despise instruments of precision. He used them in a number of his studies, but he was keenly aware that the value of an experiment cannot be measured in terms of the number or the elaborateness of the mechanical devices that are used in it. Apparatus is a means, not an end in itself, and for many types of experiments on the higher thought processes, elaborate instruments, precise measurements of simple movements, and reports of sensations must give way to careful study of the entire group of reactions to which a given situation gives rise.

By way of illustrating his point, Binet conducted an extensive and carefully planned series of studies using his two daughters as subjects. A small number of other cases were also tested in some instances. The various experiments covered a period of approximately three years, at the termination of which the older of the two sisters, Marguerite, was fourteen and a half years old, and the younger, Armande, was thirteen.

The procedures used are extraordinarily simple but they are remarkable for the keenness of the insight which is displayed in the analysis of the results. Although the report was published under the title of *L'Étude expérimentale de l'intelligence*, we should now regard it as a study of personality differences, rather than of intelligence as we are accustomed to use the term. As such it is unrivaled for the masterly way in which facts of seemingly little consequence in themselves are marshaled, one after the other, in an array that eventually leads to a remarkably illuminating analysis of the fundamental differences in the attitudes and ways of thinking of the two girls.

The exercises used were simple, but the subjects, so we are assured, took them very seriously. They were eager and interested, convinced that any carelessness on their part might have unfortunate consequences for their father's work. They were at first called upon for simple and informal tasks, such as writing a list of twenty words, the first that came to their minds. The time required for writing the list was noted, after which the entire lists were gone over with each girl separately and the

reasons why these particular words occurred to them at that time were investigated as far as the subjects were able to report them.

Certain things may be noted at this point which apply not only to this experiment but to those that follow. First, the interpretation of material such as this is obviously very difficult. Certainly it affords some evidence as to the way in which the subjects have organized their experience—which is the essence of “personality.” The question is: To what does the evidence point? That Binet was keenly aware of this difficulty as well as of the potential value of the data is apparent from his manner of treating it. He recorded the facts minutely, noted what then seemed to him to be of chief importance, and proceeded to gather more evidence. From experiments in writing words, he proceeded to similar investigations in writing short phrases, and from phrases to sentences. Always the subjects were questioned in such a way as to bring about a reinstatement, as far as possible, of the entire chain of associations that led up to the response.

From these experiments Binet passed on to exercises in memory including the reproduction of movements of the arm or hand, judgments of the passage of time, and a large number of studies of mental imagery. Again the aim of the studies was always qualitative rather than immediately quantitative, although careful records were made of the time required for completion of the different tasks and of the comparative frequency of different types of response.

At the end of his studies, Binet emerges with one of the most convincing pictures of personality differences that has ever appeared. In modern but perhaps less illuminating language than that used by Binet we should say that Armande is shown to lean rather strongly toward introversion, Marguerite to extroversion in their ways of thinking and reacting. These tendencies appear in practically all the experiments tried, but it is only by repeated studies that the fundamental principles basic in the organization of the material are brought out. In other words, Binet did not start his investigations with any preconceived notions as to what he was going to find. He was content to ascertain the facts and to follow the road along which they appeared to lead.

Some idea of the method employed is given by a brief account of the analysis of the results of one of the simplest and earliest of his studies—that on the writing at random of twenty words, followed by questioning regarding the associations which led to their choice and an analysis of the words themselves. For real appreciation of the method, Binet's discussions of this material should be read in full. Here we can do no more than present a brief outline. The data are based on sixteen trials with 20 words written on each occasion, making a total of 320

words for each of the subjects. After careful consideration, Binet came to the conclusion that the classifications of chief importance revealed by the data were the following:

1. *Unexplained words*, i.e., those for which the subject was unable to account. They seemed to occur without other associations. In Marguerite's record there were only 15 words of this class as opposed to 84 for Armande. Binet notes among other things that although separate trials with dictated material showed that the usual speed of writing of the two girls was approximately equal, during these experiments Armande wrote much more rapidly than Marguerite. Through his questioning Binet was able to show that the difference was largely attributable to what we may call "mental set." Marguerite's attention was directed, for the most part, to the sense of what she was writing. Armande tended to think of words as words, with little if any attention to their meaning.

2. *Names of objects immediately present to the senses*. In Marguerite's list, these account for 120 of the total; in Armande's list for only 30.

3. *Words referring to the self*. Binet has included here only those words having to do with physical appearance or with the subject's own clothing. Marguerite wrote 15 words to which self-reference was imputed when questioned; Armande, none.

4. *Memories*. These are words definitely referred to association with some past experience. Of these, Marguerite's list included 172; Armande's but 88.

5. *Abstractions*. Of these there are 70 in Armande's lists but only 12 were given by Marguerite.

6. *Imagination*. This classification was determined from the girls' reports of their associations, rather than from the overt character of the words written. As used by Binet, the term applies only to those cases in which a "fictive image" of some object or scene appeared to arise spontaneously in the mind of the subject and suggest the word that was written. Memories of actual experiences were not included. No such instances were ever reported by Marguerite, but Armande's total is 23 unquestionable cases in addition to others in which the presence of the mental image is less certain.

The combination of the basic principles of the "projective" approach with wholly objective methods of treating the material which is shown in this study may well serve as a model for present-day psychologists who recognize the possibilities of an indirect approach to the more subtle aspects of the personality but have lacked the insight to apply the methods of exact science to their data. No one was more



FIG. 34. BINET'S TWO DAUGHTERS. (Courtesy of Lewis M. Terman.)

keenly aware than Binet of the dangers of hasty judgments, of generalizations based upon surface impressions or upon single occurrences. All of the differences in the foregoing analyses would meet even the most rigid of modern requirements for "statistical significance," yet this series constitutes only a small part of the experiment. Compared with it, most of the modern projective methods appear superficial. We turn now to a consideration of a few of these methods.

DOLL PLAY

In the search for causes of maladjustment among children and adults, family relationships and particularly the relationships between parents and children have been subjected to close scrutiny. Social conventions, however, have all but completely forbidden many of the more usual ways of getting at the problem. Both by direct and by indirect means, most children are indoctrinated at a very early age with the idea that members of a family are supposed to love each other and to love each other equally, that hostile feelings must not be put into words or family affairs discussed before strangers. Although these attitudes may be partially broken down by the sympathetic interviewer, it is still not easy for the child to put into words what he has never felt free to verbalize, even to himself.

The use of dolls designed to represent the various members of a child's family is a very popular way of studying this problem. Originally the method was used only with young children. Some people, however, have extended the device to the ages up to and including adolescence and even to the adult level. Older subjects are usually instructed to make up a play in which the dolls are to serve as puppets. For this purpose a toy stage is often provided.

With younger subjects the dolls are often identified, with the child's help, as members of his immediate family. There are Daddy and Mother, the baby, perhaps an older sister or brother, and the child himself. The child is then told, "Now I have some writing to do, so I will sit over here at this table while you play with the dolls. You may do whatever you wish with them. It doesn't matter if you break them."

In addition to the dolls, toy furniture is frequently provided. This usually includes a set of bathroom equipment and a bed. Sometimes other articles are included but it is noteworthy that such items as toy dishes and kitchen and dining room furniture are rarely mentioned.

Detailed records of the child's play are made, and all verbal comments are recorded verbatim. Records of suggestions and other comments made to the child are sometimes kept as well, but as a rule this part of the pub-

lished reports is rather sketchily handled. Many examiners, however, appear to make a practice of inciting the child to handle the dolls roughly by such remarks as, "Don't be afraid of hurting them. You may do whatever you like. I don't care if you break them." (Baruch, 1940).

The assumption made in all the studies involving doll play, puppets, or similar material is that the child identifies the dolls with certain persons, usually members of his family, and either treats them as he would like to treat the real persons whom they symbolize, or through imitative play dramatizes his conceptions of their relations to each other and to himself. Such an assumption implies that attitudes of this kind lie very near the surface and will be revealed in terms of symbols that are but thinly disguised whenever opportunity is given. This may be true, particularly in the case of young children upon whom convention has as yet made but slight impression. But it is possible that some of the "hostile" reactions, so vividly described by Baruch (1940), Despert (1940), and many others, may be inspired by nothing more than the delight children usually take in the form of construction that adults are likely to regard as destruction,¹ and thus have little or no symbolic meaning for them. When the only toy furniture provided consists of a toilet, a bed, and an armchair, one should not be surprised if much of the child's play centers around the first two. It is questionable whether such play can be looked upon as a valid sign of the child's interest in sexual matters. What the choice of materials and the interpretation of the child's play may indicate with respect to the personality and interests of the examiner is another matter which will be considered in a later section of this chapter.

THE INTERPRETATION OF ART PRODUCTS

The belief that the drawings and paintings of mentally disturbed persons may help to reveal some of the factors that are causing their distress is not new. Early in the present century a number of reports based chiefly upon single cases were published. These reports tended to show that in drawings by persons showing marked obsessions with some single topic, such as religion or sex, some particular symbol or symbols tended to occur and recur. Sometimes these symbols were unmistakable. Some-

¹ For the young child who is not yet able to do much in the way of putting things together, taking them apart is a really constructive activity. He is making something that was not there before, something that has more pieces and perhaps greater possibilities of manipulation. If the process involves interesting noises and requires such amusing activities as banging, hammering, or trampling, so much the better. And finally, if, in his small way, he is aware that what he is doing is usually forbidden, the situation takes on the added fillip of daring and adventure.

times they were drawn first and then concealed under other features so that only by close examination could they be identified. Sometimes they bore no overt resemblance to that which they signified for the subject. Their meaning was learned only by careful questioning or by inferences based on the subject's general behavior and incidental remarks.

By far the most careful study of the drawings of mentally diseased persons is that by Prinzhorn (1922), who collected many hundreds of drawings and paintings made by patients in mental hospitals, and noted the features that seemed most characteristic of the art products of each person. He then selected a small number of patients, most of whom were of the schizophrenic type, for special study. All of these subjects spent a good deal of time in drawing and painting so that it was possible for the investigator to study a large number of examples of the work of each one, and to note the recurrent features of their style and subject matter. Some of these had to do with the manifest content of the patient's mental life; they were attempts at representing the things that he chiefly talked about. They had to do with his overt fears, wishes, anxieties, or preoccupations. Others were covert symbols; they stood for those things of which the patient was unwilling or perhaps unable to speak.

During the past decade, the use of drawing as an aid to psychological and psychiatric diagnosis of personality difficulties has been steadily gaining in popularity. Unfortunately, most of this work, like that of play analysis, has been carried out on a decidedly superficial level. The assumption commonly made is that the subject will reveal his difficulties in bald outline once he is given pencil and paper; that he will not hesitate to express graphically that which no ordinary means will induce him to express verbally. Great significance is therefore attached to drawings which in all probability represent nothing more than a passing interest on the part of the child. I recently attended a solemn gathering of nursery school teachers who had met with a number of psychiatrists and psychologists, to discuss the serious problem presented by Tommy, age four, who, in addition to other misdemeanors had recently painted a large black blob which he proudly designated as a "wolf." The painting was handed round and examined with much headshaking. Tommy, it appeared, was interested in wolves. Presumably he admired wolves. Therefore he would like to be a wolf. The final step was quickly taken. *Tommy believed himself to be a wolf.* Wolves are ferocious animals. Tommy must therefore be ferocious. In his fight of that day with a boy considerably larger than himself he undoubtedly was taking the part of a wolf. Tommy's problem was now diagnosed; it remained only to find a solution. Curiously enough, no one seemed to think of teaching Tommy to paint rabbits and lambs instead of wolves!

It is easy to be facetious about this sort of thing, but as a matter of fact no one, perhaps, is more firmly convinced than I of the *potentialities* of drawing as an area in which to look for signs both of the personality characteristics of children in general and of the particular difficulties which affect individual children.² I question, however, that the road to identifying these signs and to interpreting their meaning is as simple and obvious as most of those who have attempted to use the method have assumed. I question the existence of a one-to-one relationship between the surface aspects of a child's drawing, particularly as seen and interpreted by an adult, and his private world of feelings and emotions, many of which he may be unwilling to admit, even to himself. The identification of a symbol with the thing overtly signified is a basic feature of the "sympathetic magic" of primitive man. Has an analogous principle crept into some of our psychological concepts?

Even more strongly do I question the "interpretations" of drawings made on the basis of adult theories as to what they might mean but with little or no evidence as to what they do mean. Wolff (1946) has given us some illuminating examples of the manifold ways in which children translate their feelings and impressions into graphic form, such, for example, as occur when a piece of music is played to a group of young children who are then asked to "draw the music." But he is on insecure ground when he attempts to assign definite symbolic meaning to particular features of individual drawings, even when the interpretation is made in part on the basis of the child's comments during the act of drawing. The procedure seems to be pretty much as follows. Starting with certain basic premises concerning children's feelings and attitudes,³

² The reason for specifying children's drawings, rather than those of adults, lies in the fact that drawing is much more commonly used as a means of expression by them. Nearly all children like to draw. Few adults are accustomed to do so, and they are likely to display considerable reluctance if urged to draw in the course of a psychiatric interview or a psychological examination. It may be noted, however, that Berdie (1945) found that the Goodenough "Draw a Man" test worked fairly well as a rough screening device for the identification of mentally defective naval recruits. More recently John N. Buck has been at work on a test intended to reveal certain personality characteristics of adults which is based on an analysis of their drawings of a house, a tree, and a person. After the drawings have been made, the subject is questioned about them according to a formal plan devised by Buck. Preliminary reports on this device look promising. Another line of approach that might have possibilities is to be found in the "doodlings"—the scribbled sketches and casual designs that many people make during periods of boredom or when using the telephone. A few informal studies of these half-unconscious scribbles have been reported both in the psychological literature and in popular magazines and newspapers, but the method used has not been such as to provide much information of psychological value.

³ In Wolff's case, a premise upon which he lays particular stress is the animosity toward a new brother or sister which he assumes to be well-nigh a universal feeling among children. He therefore seems to take it for granted that this feeling will be

Wolff searches for evidences of his beliefs in the drawings of the children whom he studies. The reports have a certain plausibility, owing in part to the fact that only evidence in line with the explanations offered is given and no alternative explanations are suggested.

Alschuler and Hattwick (1947) studied the drawings and paintings of 150 preschool children over a period of one year. Twenty of these children were also followed during a second year. Although many aspects of child art are considered in the two large volumes of their report, their particular concern is with the significance of color preferences and of the manner in which colors are applied. Both case studies and statistics based on the work of the group as a whole are presented. The case studies are internally consistent and the interpretations seem in line with what is known of child psychology. Although few, if any, of the reported differences among selected groups are large enough to meet the usual criteria for statistical significance, at least the facts are given in a manner that permits evaluation.

Both Wolff and Alschuler and Hattwick offer many suggestions for further work in this promising field. Whatever criticisms may be offered

symbolized in the drawings of children and examines them carefully for features that may be interpreted in this way. For example, there is the case of Ellen (pp. 113 *et seq.*), who drew pictures of cats which are interpreted as symbols of herself. During the soliloquy which accompanied one drawing, twenty-one separate items were enumerated as she drew them. Of these, eight are described as "beautiful," "nice," etc., while twelve are either merely named without description or described as "little," "new," "black" (referring to a shoe) or in other nonevaluative terms. But when she came to the claws, Wolff states that "an emotional expression came over her face" as she stated, "And here are his sharp claws." Granting that Wolff's interpretation of the child's facial expression may have been correct, is one warranted in the assumption that this drawing symbolizes the child's feeling of aggression, or her wish to display aggression toward the new baby in the household or possibly toward her parents? Is it not at least equally possible that any emotion associated with the pictured claws may have arisen from a remembered personal contact with the actual sharp claws of a real cat? (It is noted that Ellen had at home a cat, which had young kittens.)

A second example, again taken from Wolff (p. 128), is that of the "birth fantasy" of a four-year-old boy whose mother had not told him that she was expecting a new baby. There is no evidence that the child showed any special curiosity about the baby's coming; as a matter of fact, it is specifically stated that he asked no questions about it. But, Wolff adds, "the pictures are such questions." However, the mere fact that the child drew a number of pictures showing chickens coming out of eggs is to me a very inadequate basis for such a statement. It has been repeatedly shown that children rarely see any relationship between the hatching of eggs and the birth of a baby.

On one occasion, this boy drew a picture which he described as "My great big chicken; it's bigger than the nursery school." Wolff commented that the picture "looks like a totem pole" and offered the following account of its meaning: "As the child did not get an answer from his mother [to what is not stated; it will be recalled that the child did not ask questions concerning birth] he prays to his totem animal because this animal, laying eggs, should know the answer to the secret of birth." No eggs, incidentally, appear in the drawing, which Wolff reproduces on p. 129. Eggs, prayer, and question—all seem to have been supplied by the investigator.

of their reports, it is difficult to read them without becoming convinced that here is an area which should richly repay further study by more adequate methods than have been used up to the present time.

OTHER METHODS USED PRINCIPALLY WITH CHILDREN

Many kinds of play situations making use of a wide variety of materials have been set up, and the behavior of children in these situations has been analyzed according to projective theories. As in the case of drawings, interpretations have for the most part been rather shallow, based chiefly upon some type of surface manifestations with little or no attempt at verifying the conclusions reached. In addition to the dolls previously mentioned, fragile objects of various kinds, notably rubber balloons, which the child is encouraged to break have been used as a means of studying "aggression." Smeary substances, among which cold cream appears to be a favorite, have also been used, the theory here being that the child who daubs himself thoroughly thereby gives evidence of anal eroticism or conflict over some aspect of sex. Both these ideas seem a bit far-fetched, and little in the way of supporting evidence has been offered except for descriptions of a small number of individual cases which may or may not conform to the general rule. Other devices used both with adults and with children have been the psychodrama, supplying an ending to stories, constructive work of various kinds, and many other devices of a roughly similar nature. In all this, the major difficulty seems to lie in the fact that most of the experimenters have been too eager to secure striking results immediately. They have jumped at conclusions from flimsy evidence. They have neglected the most elementary rules having to do with such matters as sampling, agreement with objective criteria, and other basic principles of scientific procedure. These criticisms, needless to say, do not apply to all. A small number of persons are making a serious attempt to clear away the rubbish with which the field is now cluttered and expose the fertile soil that lies beneath. To them we may look for methods that will last.

METHODS USED CHIEFLY WITH ADOLESCENTS AND ADULTS

FREE WORD ASSOCIATION

The use of free word association as a means of uncovering emotional conflicts has been a well-recognized technique in psychiatric interviewing since the days of Jung. Freud also used it, though less extensively than

Jung, for whom it became a well-nigh indispensable tool. In the psychological laboratory many experiments in the use of word association have been carried out, and a number of diagnostic signs by which the particular words that impinge upon the "sore spots" in the subject's conscious or unconscious mind may be identified. Among these signs are unusually long delays before responding, evidences of embarrassment such as flushing or giggling, changes in tonal inflection, and the like. By using a telegraph key attached to a kymograph, Burt (1936) found that when the subject was told to depress the key at the same time that he gave his response, the kymograph record revealed other signs which were shown to have diagnostic significance. (See Figure 35.)

More recently Goodenough (1942, 1946) has been able to show that personality factors of a more general kind, such as mental masculinity and femininity, leadership, and communality of response (similarity to the responses given by others of corresponding age and sex), can be identified by the application of statistical methods to the free associations to common words given by large groups of subjects. For this purpose, Goodenough made use of homographs—words which, in their written form,⁴ have two or more different derivations and therefore two or more distinct meanings. Scoring keys for this test have been devised but have not yet been published. Self-correlations for groups of like sex and fairly homogeneous age range⁵ are in the neighborhood of $+.85$ to $+.90$ for each of the various keys thus far derived. Maurer, who made a special study of the leadership key (1947), found that by its use college women who took a leading part in many campus activities were differentiated from nonleaders with very little overlapping between the two groups. The method should repay further study.

MURRAY'S THEMATIC APPERCEPTION TEST

This test, which is based upon the interpretation of pictures, is widely used in psychological and psychiatric clinics because of the many leads it provides for the understanding of the individual personality. From the complete series of thirty-one pictures, special sets are made up for use with younger or older subjects of each sex. In the test, the subject is shown the pictures in a standardized order and is asked to make up a short story about each one. A number of different scoring methods have been proposed, none of which are highly objective since the wide

⁴ Goodenough used a written form of the association test, in which the stimulus words were printed on the blank with space provided for the subject to write his response. The group-testing method was used.

⁵ Not greater than two years for the ages below twenty; not over ten years for adults.

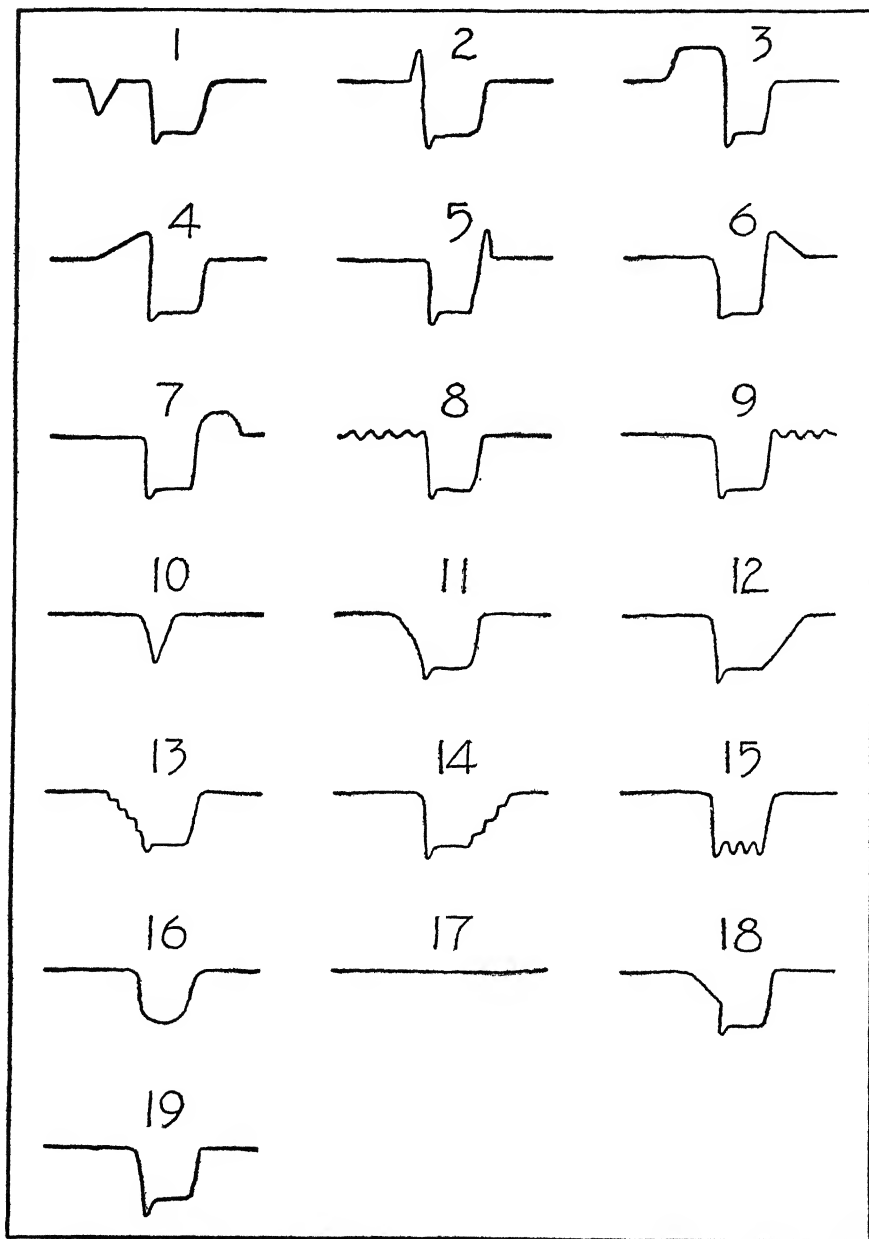


FIG. 35. DIAGNOSTIC SIGNS IN THE MOTOR RESPONSE TO WORDS HAVING EMOTIONAL SIGNIFICANCE FOR THE SUBJECT. (After H. E. Burtt, "Motor Concomitants of the Association Reaction," *Journal of Experimental Psychology*, 1936, 19: 51-63.)

individual variations among the responses render a completely formalized scheme difficult of accomplishment. Furthermore, a wholly objective scoring method would do away with much of the value of the test because it could make no adequate provision for the unusual or "individual" interpretations which are frequently the ones of chief significance.

Features to be especially noted are indicated in the author's *Manual* (1943). Among these are (1) the pictured character with whom the subject apparently identifies himself; (2) the kind of external forces that occasion the hero's behavior or give rise to the emotions which he is alleged to be experiencing (to these, the author gives the name of "press"); (3) the nature of the outcome, whether happy or unhappy, tragic, farcical, and so on.

As commonly used, the TAT is an individual test. Clark (1944) attempted to use it as a group test in which the pictures were projected on a screen. Fairly high correlations were found between two ways of administering the test. In the first, called the "clinical test," the subjects were told to write a short story about each of the pictures as it was shown. They were told to be sure to state what they thought had happened to give rise to the situation shown (press), what the subjects were thinking or feeling, and the outcome. In the second test, given two weeks later, the same pictures were used, but instead of making up their own stories, the subjects were given a list of alternative descriptions from which they were to choose the one they thought most appropriate. Whether the generally high agreement found between the results of the two methods indicates a real stability of the manner in which the different subjects are accustomed to interpret the behavior of others, and hence, to some extent, of the diagnostic value of the test, or whether this agreement can be largely or wholly accounted for on the basis of memory cannot be determined from the data presented. Some use of the TAT as a group test, using a similar method, was also attempted in the army during World War II.

A number of persons have applied the general method of the TAT to groups whose level of development or social background made the standard pictures so inappropriate that others were devised. Temple and Amen (1944) worked out a series for young children so planned as to elicit indications of anxiety. Fairly marked differences were found in the percentages of "happy" interpretations given by a group of seven children in an institution where most of the cases came from homes broken by death or divorce, and by those from an ordinary nursery school and those from an apparently well-run sanitarium for cardiac cases. Individual reports show considerable surface agreement between the general type of picture-interpretation given by the child and the known facts in his

experience, but, as in other instances of this kind, there is no way of determining the existence or nonexistence of bias in the selection of these cases.

Henry (1947) had an Indian artist make a series of drawings depicting various scenes in American Indian life. These pictures were shown to Indian children who were asked to make up stories about them. Responses were scored according to the TAT method. In summarizing his findings, the investigator affirms that the procedure is unquestionably useful as a means of studying the more intimate aspects of the life of societies other than our own and particularly those that have to do with the home and family. Descriptions of the kind of differences between Hopi and Navaho groups suggested by these records were found to be in general agreement with anthropological reports and with the opinions of anthropologists who read the descriptions. Henry calls attention, however, to the great difficulty of evaluating this material in such a way that it may justifiably be transformed into reasonably objective data on the personality characteristics of the subjects who provide it.

As is the case with the Rorschach method next to be described, two methods of handling the TAT material, each based upon a somewhat different concept of the nature and organization of personality characteristics, are in common use. The first is the method of "blind analysis," favored by those who insist that the personality of each individual is an indivisible whole, that attempts at statistical analysis into "traits" can only destroy the fundamental harmony which is its essential feature. They stress a point mentioned here before: that a characteristic which makes for beauty and strength in one connection is a source of ugliness and strain in another; that "good" and "bad," "desirable" or "undesirable," "wise" or "foolish" are terms that cannot be used as absolute descriptions with constant reference but must always be considered in terms of the total situation. This is the so-called "holistic" view of personality.

In the method of blind analysis, two descriptive sketches of each individual are used. The first is prepared by the person who gives the test and is based entirely upon the impression of the individual thereby gained. The second is made by a person who has no knowledge of the test results but who is intimately acquainted with the subject. Sometimes a general guide to be followed in drawing up the descriptions is provided; in other cases the writers are left completely free to note the points which they regard as most significant.

The two sets of descriptions are then presented in random order to one or more persons who are not acquainted either with the subjects or with their test performances. The problem is to see how accurately

the pairs can be matched. Most studies of this kind that have been carried out with the TAT have shown better than chance agreement, a fact which indicates that the method has some validity for the analysis and description of personality (Harrison, 1940).

The second method involves counting the number of responses of certain types given for all the pictures in the series and using the totals as scores⁶ from which separate aspects of personality are inferred. This procedure is based on the analytic or "trait" concept of personality. When used, self-correlations and correlations with other criteria are determined in the usual fashion. Regardless of which method is employed, users of this test have relied mainly upon case history for the demonstration of its usefulness.

THE RORSCHACH INKBLOT TEST

Few, if any, psychodiagnostic methods have leaped into prominence with the velocity exhibited by the Rorschach test. Rorschach's original monograph describing the method was first published in Switzerland in 1921. As was true of Binet, Hermann Rorschach died before his method had been perfected, and his work had to be carried on by others. His brilliant pupil Emil Oberholzer, who took the leading role, not only was responsible for the first publication describing the method to appear in the United States (Rorschach and Oberholzer, 1924) but trained the American psychologists who first introduced it into this country. David Levy is generally credited with being the first to introduce the method here, and it was his influence that sent Samuel J. Beck to study with Oberholzer and later to write the first American doctoral dissertation on it. In 1937 Beck published the first complete manual for the use of the Rorschach to appear in English.

The Rorschach procedure is not easy to master. Many who wish to become expert in its use are daunted by the exceedingly elaborate and detailed analysis required, or are dismayed by the fact that, unlike the other projective methods which have been described, interpretations cannot be made on a merely superficial level. The Rorschach worker does not look for surface resemblances between what is described and the individual personality. The signs for which he seeks go deeper than that.

⁶ Weights are sometimes assigned to the responses given to the different pictures on the basis of the infrequency of a given type of response. The assumption here is that if, for example, a subject ascribes a sorrowful meaning to a picture which to most people suggests happiness or content, the inference of unhappiness or anxiety on the part of the subject is stronger than would be the case had the picture been similarly interpreted by the majority.

The difficulty of learning to administer and interpret a Rorschach test led, in the early days, to many discrepancies in the methods used by amateurs attracted by the novelty of the procedure and the apparently spectacular results achieved by it. For a time the method seemed to be on its way to becoming a fashionable cult rather than a serious attempt to learn the hidden meanings of a form of expression through which the springs of human behavior might be laid bare. To some extent, this situation still exists, but it has been in part corrected by the establishment of better facilities for the training of workers⁷ and by the increasing interest in the subject displayed by clinical and research psychologists of established reputation. Such recognition is gradually crowding out the charlatans.

The Rorschach material consists of ten standardized inkblots, similar to but not identical with that shown in Figure 36. Five of the blots are in black and white only; the other five include color. Each blot is printed on a separate card, 7 by 9½ inches in size, numbered in the order in which they are to be shown. There seems to be no well-established form of instruction to the subject since almost any one that may be chosen involves, as Klopfer and Kelley (1942) have pointed out, the possibility of biasing his response. In general, however, after some preliminary explanation designed to put the subject at his ease and also to make the purely accidental character of the blot clear to him⁸ so that he will not be concerned about the possible existence of some concealed form that he is expected to see, he is given some such general instruction as "Look at this blot and tell me what you see—what it makes you think of." Questions are answered in noncommittal terms such as, "Whatever you think," or "That is all up to you."

The remaining blots are shown in the same way. Responses are scored in terms of four general categories for which exact instructions are given in the manual by Klopfer and Kelley (1942). These are (1) the *location*, that is, the part of the blot to which the response refers; (2) the *determinant* of the response—such as color, shading, or form; (3) the *content*, that is, the kind of object named; and (4) whether or not the content is one frequently perceived by others (a "popular"

⁷ Such, for example, as the creation of summer workshops under the direction of Rorschach experts; courses in some of the leading colleges and universities, the publication of the *Rorschach Exchange*, a quarterly journal devoted entirely to reports of research on the method; and the setting up of the Rorschach Institute for the training of workers in the field and as a center where advice on knotty points of administering and scoring the test can be had.

⁸ Some examiners demonstrate the method of making the blots by dropping a few sprinkles of ink on a piece of paper, folding and pressing it to spread the ink and so make a picture similar to those shown.

response) or an unusual or "original" response. This is determined by reference to a table based on a tabulation of the responses given by a large number of subjects previously examined. Time required for responding is also recorded.

Interpretation of a Rorschach record is based upon two things. First there is the formal scoring in terms of rather careful instructions



FIG. 36. AN INKBLOT SIMILAR TO BUT NOT IDENTICAL WITH THOSE USED IN THE RORSCHACH TEST.

for each of the categories mentioned above or, more precisely speaking, for a fairly large number of subheadings under each. This scoring is reasonably objective, and the self-correlations of the various items have been found to be rather surprisingly high by practically all who have studied the matter. Second, the trained Rorschach worker usually adds a subjective interpretation, based on the record as a whole. The validity of these subjective impressions, when made by well-trained and experienced workers, has been repeatedly checked by the method of blind analysis described in the preceding section. Agreement is usually well beyond chance expectation.

In the belief that the interrelationships of the scores on the various items of a Rorschach record may be as significant as the scores themselves, a number of attempts have been made to derive "formulas" by which certain arithmetical combinations of these scores are made and the result is used as a sign of some personality trait. This work is still in the experimental stage.

In comparing the Rorschach method with other projective techniques, a number of important differences may be noted. There is no obvious resemblance between the nature of the subject's responses and the personality traits from which they are alleged to spring. It is necessary to learn the meaning of each response by a slow and laborious comparison of the characteristics of those who give that response with those of others who do not, or, conversely, by a comparison of the responses of groups who are known to differ in respect to some reasonably objective aspect of personality or behavior. Although it is unfortunately true that a good many of the "signs" reported have stemmed for the most part from "hunches" with but small basis in the way of experimental evidence, theoretically, at least, the Rorschach signs are interpreted by statistical rather than by intuitive processes.

Certain questions with regard to the method may be raised. Since only ten blots are used, the signs from which interpretations are made are often few in number. Although alternative series have been tried, their use has not been widespread; most workers confine themselves to the original set of ten blots. There are two reasons for this: first, the very great amount of work required for the adequate standardization of additional material; and second, the fact that, as usually given, the administration of the test in its present form requires about as much time as is likely to be available.⁹ Nevertheless, it is possible (though by no means certain) that the addition of more blots and the elimination of some of the less objective features of administration and scoring or possibly the introduction of a time limit would increase the dependability of the record without loss in its diagnostic value.

Because the differentiating signs are not immediately obvious, either in respect to their surface characteristics or in respect to the deeper meaning they signify, it was perhaps natural enough that a large number of the studies made up to the present time have been exploratory rather than thoroughgoing. Signs have been pointed out with enthusiasm from

⁹ When the test is given to an individual, no time limit is usually imposed, and the subject is encouraged to continue responding as long as he is able, turning the card to different positions as he wishes. The length of time required is further increased by the "inquiry" at the end, in which the subject is asked to explain in greater detail those of his responses of which the nature or origin is not clear to the examiner.

evidence that has later proved to be only circumstantial; "formulas" have been developed on the basis of chance relationships noted for a small number of cases, with little or no attempt to apply ordinary methods of statistical analysis to ascertain the level of confidence that may be placed in the results. All this is not unusual when enthusiasm for a new and promising method runs high. It was true in the case of intelligence testing. But the Rorschach method has now reached a point of development where a more critical attitude is desirable. That it is capable of yielding information of considerable value for the study of the individual seems reasonably certain, but much winnowing is still needed to separate the wheat from the chaff.

The Harrower-Erickson group form of the Rorschach (1945) was much used in the army during World War II as a screening device; it has also been used in schools and colleges both for research purposes and as a clinical method for selecting students who may be in need of special advice or guidance. The blots are projected on a screen by means of slides. The subjects are provided with test blanks on which answers are to be checked from multiple-choice lists. No claim is made that this method will indicate the fine shades of diagnostic meaning provided by the individual method. Its advantages are the obvious ones of timesaving and the substitution of a formal device that can be used by persons with but little training and experience for the intricate analysis that can be made only by a highly trained expert. Within its limits the group test has been shown to have a certain value, but it should not be regarded as an adequate substitute for the individual Rorschach.

OTHER DIAGNOSTIC METHODS BASED UPON THE THEORY OF PROJECTION

That the personality of an individual is to some extent revealed in his handwriting is a widely accepted belief. The scientific analysis of handwriting has, however, been handicapped by a number of factors. Handwriting styles change with time and are much affected by school training. Exploitation of graphological methods by charlatans has been all too common and has tended to bring the entire method into disrepute in the eyes of many persons. Nevertheless, the mere fact that the sex of the writer can be correctly judged from a single specimen of his handwriting in approximately two of three cases chosen at random¹⁰ is sufficient proof that "something" is there. Here as in other projective methods the great difficulty lies in interpretation, and few indeed are the investigators who have resisted the temptation to arrive at quick and

¹⁰ See Downey (1910), Newhall (1926), Goodenough (1945), and others.

easy generalizations in terms of surface impressions and popular symbolism. An excellent review and critique of the experimental literature in the field of graphology is to be found in Bell's book on projective methods (1948).

A possible means of validating graphological analyses is to be found in the use of specimens of the handwriting of famous men about whose achievements and personality characteristics a good deal of information is available. Although a number of such studies have appeared in the literature, the most recent of which is by Wolff (1948), the treatment of the data has been for the most part impressionistic. Little use has been made of statistical methods for the selection of signs and the determination of probability levels. In this connection it may be well to call attention to the desirability of distinguishing between positive and negative errors when the signs derived from one group of subjects are checked by means of a second group used for validation. "False positives" include those cases in which the sign presumably signifying a given characteristic is present in the handwriting or other type of performance of an individual who does not exhibit the characteristic in question.¹¹ "False negatives" are those cases in which the sign fails to appear when the characteristic is present. As a rule, the former type of error is likely to have the more serious consequences, but this is not invariably true.

The Tautophone, which is an adaptation by Shakow (1940) of Skinner's Verbal Summator Test (1936), is another example of the use of a slightly structured situation which the subject can complete as he will. The device consists of a set of phonograph records played at low intensity on which a series of vowel sounds are repeated in irregular order in such a way as to give an illusion of human speech. The subject is told to listen and report what is being said as soon as he can understand it. An elaborate system of scoring has been worked out, but upon the whole the method has not proved as satisfactory as had been hoped. In part the difficulties arise from the wide variety of responses given, which vastly aggravates the problems of classification and scoring. It is also possible that too much has been expected from a device that is still in its infancy. Thus far the Tautophone has been used for the most part in mental hospitals as a means of helping to differentiate the various forms of mental disorders. The criterion for such distinctions is not wholly objective and the overlapping of symptoms is such that an instrument of considerable delicacy is required if the task is to be per-

¹¹ In the case of continuous characteristics which all possess to a greater or less degree, presence or absence is interpreted to mean a level that falls above or below the average of the group to which the subject belongs.

formed satisfactorily. Perhaps more could be accomplished by beginning with broader classifications and deferring the more difficult problems until a greater refinement of the procedure has been worked out.

The belief that character can be judged from physical appearance is implicit in the common practice of requiring a photograph as a part of the application for various types of positions. Many experiments have been conducted to test the validity of this hypothesis. In practically all cases, however, the correlations obtained between the judgments of personality characteristics made on the basis of the photographs and the results of objective tests or ratings of the subjects themselves have been little if at all better than chance. A modern adaptation of the method is the Szondi test of which an American translation has been published by Deri (1948). This test consists of six sets of eight photographs each. The photographs depict a number of abnormal types—a homosexual, an epileptic, a depressive, a hysteric, and so on. All the pictures in a single set are shown to the subject simultaneously. He is told to select the two liked best and the two liked least. The assumption here is that an unconscious process of identification will lead to a preference for those characters that have something in common with himself. It should be noted, however, that even if the hypothesis were true—and its truth has never been satisfactorily demonstrated—a number of difficulties remain. For each individual pictured must have possessed not merely one, but many characteristics. A homosexual person, for example, is not just a homosexual and nothing else. He is also a carpenter, a salesman, or a member of some other occupational group. He is intelligent or stupid, gregarious or solitary, healthy or sickly, and so on through a well-nigh endless list of attributes. If identification of the kind assumed by the Szondi test does occur, there seems to be no particular reason why it may not be made upon the basis of any of these characteristics, rather than on the single one on the basis of which the selection was made. Although the history of physiognomy as a method of psychological research has not been such as to inspire optimism, it may be noted that the work of Wolff (1943) does provide some evidence that may perhaps be construed in favor of Szondi's theory. Wolff was able to show that the responses of many persons toward photographed reproductions of parts of their own bodies, or their own behavior when these were presented in such a way that overt recognition did not occur, nevertheless differed in a number of characteristic ways from those made to similar material obtained from other persons.

The Rosenzweig Picture Frustration Test (1945) is allied both to the Thematic Apperception Test and to the various tests of story completion. It consists of a series of cartoons. Each cartoon depicts two per-

sons, one of whom is saying or doing something to annoy or "frustrate" the other. The subject is asked what he thinks the second person would do or say under such circumstances. Scoring is done in accordance with a prepared scheme that takes account both of the character of the response and of the object against which it is directed.

The World Test designed by Charlotte Bühler (1941) consists of a large number of small figures mounted on wooden blocks so that they will stand upright. The figures include human beings, buildings of various kinds, animals, trees, and sections of a wooden fence. The subject is told to arrange them on a wooden tray so as to construct whatever kind of scene he wishes. These scenes are scored in terms of a number of descriptive categories designated as "worlds." Such matters as the number of pieces used, the time spent in building, the number of changes or corrections made, and the like, are also noted. The following are examples of the several kinds of "worlds" thought to have diagnostic significance: (1) the "closed world," which is entirely surrounded by a fence and is said to characterize a timid, anxious type of personality; (2) the "rigid world," where all the objects are disposed in precise rows or patterns, a plan thought to indicate the presence of many inner conflicts and inhibitions; and (3) the "disorganized world," in which the objects appear to be placed in helter-skelter fashion with no observable plan of organization. This is thought to be diagnostic of a state of doubt and confusion which is in many cases attributable to a feeling of insecurity. A number of other types have been pointed out. Although the test was originally designed for use with children, it has also been used with adults. Except for a few case reports, very little evidence of its validity is available.

This brief account does not, by any means, cover the entire range of projective methods that have been devised and more or less completely standardized. Although their forms differ, the main problem with which all are confronted is the same. This is the question of meaning, a question rendered the more difficult of solution because of the extremely wide range of responses permitted by most of the methods used. A further complication arises from the possibility that the symbolic languages of different persons may not correspond, that responses may vary in meaning according to the person who makes them and the circumstances under which they are made. All this shows the need for tempering enthusiasm with caution.

As is true with all diagnostic methods, case reports provide useful illustrative material but they must not be looked upon as substitutes for statistical evidence. Among the facts that should be—but too often are not—reported for tests of this kind are the following: (1) the extent

of agreement among observers in classifying a given set of responses, (2) the consistency of the responses of the same subjects on different occasions, (3) the validity of the conventional interpretations as indicated by the correlation of the test results with other evidences of the personality of the subjects, (4) the uniformity of the above relationships for subjects of different ages, sex, and experiential background. In the absence of such information, a healthy degree of skepticism with respect to the usefulness of many of the new and half-tried "projective methods" with which the market is being flooded may well be maintained.

CRITICAL DISCUSSION OF THE PROJECTIVE THEORY AND METHODOLOGY

At the beginning of this chapter it was noted that the fundamental theory underlying all projective techniques is that every individual tends to project his own feelings and attitudes upon the objects and people in the world by which he is surrounded. His perceptions of this world are thus determined only in part by its physical character; in part they are reflections of himself.

The assumption is therefore made that in the individual peculiarities of these perceptions either as verbally reported or as indicated by such forms of graphic expression as drawing or painting, literary compositions, dramatizations, and the like may be found clues from which the structure of the individual personality may be discovered. That which cannot be seen directly may nevertheless be observed, so it is assumed, in the form of its image as projected upon the outer world.

A point that is often lost sight of, however, is the fact that projection, as thus conceived, is a tool that cuts both ways. Not only does the person who is observed project his own inner feelings and attitudes upon the situation to which he responds, but the person who observes him does the same. Psychologists are human beings. As such, they are by no means exempt from the laws that govern their kind. So the psychological examiner projects his own personality upon the subject he observes in the course of a "projective experiment." He views the subject's activities and the products thereof in the light of his own preconceived beliefs, his own feelings and attitudes. He interprets them in the language of his own private world.¹²

¹² Some years ago I had occasion to make use of a large number of social case histories prepared by a group of about eight or ten psychiatric social workers. I soon found that long before I came to the signature at the end of a record I could identify the worker who had prepared it with almost complete certainty, even though all the records had been typed by the same stenographer in uniform style. Not only were there individual peculiarities in literary style, but even more pronounced differences

It is largely because of the fact that many of the projective methods provide almost complete opportunity for the examiner to project his own beliefs and attitudes upon the phenomena that he perceives, and to inject his reactions to these perceptions into his conclusions and diagnoses, that the methods have been regarded with so much skepticism by many of the more "hard-boiled" experimentalists and clinicians. To these persons, the case histories upon which workers in this field have laid so much stress may provide useful illustrative material with which to drive home a principle or enliven a discussion, but because they have been *selected* in ways that are rarely described there is no assurance that they represent a rule rather than an exception to that rule.

If projective methods are to win the unqualified approval of scientific workers, it is necessary to abandon the attitude that scientific proof can be dispensed with in favor of intuitive judgments often based upon a kind of evidence that smacks dangerously of the witch doctor who mutilates an image to injure his enemy. Selected case histories do not constitute proof; neither do the pages of little stories and anecdotes by which a surface appearance of plausibility is sometimes attained. The way of science is not the easy road of surface impressions and pronouncements by fiat. It is a toilsome road, with many rough places that only arduous labor can render passable.

appeared in the content, in the kind of things that had been noted and thought worthy of recording. To some, neatness was of paramount importance. A crumpled blouse, a broken shoelace, an unwashed plate in the sink, or a newspaper lying in a chair could not be overlooked or excused. Others were preoccupied with ascertaining the client's attitude toward himself. "Inferiority feelings" were just then all the rage, and indications of such feelings were unfailingly detected by these workers. Still others found moral delinquency lurking on all sides; others were equally certain to discover indications of malnutrition or of mental backwardness. On the whole, the histories appeared to provide about as much information concerning the persons who prepared them as was afforded concerning the subjects to whom they presumably referred.

Tests for Vocational Guidance

THE AIMS OF VOCATIONAL GUIDANCE

Guidance, as the name implies, consists in helping someone find his way to a goal that is still ahead of him. In vocational guidance, the goal to be sought is a useful and satisfying field of work. The help to be given has to do mainly with the selection from the many fields that meet the criterion of social usefulness; one that also meets the second criterion—that it shall be satisfying to the individual. The latter criterion implies that the field chosen is one in which the individual can meet with success. Properly speaking, vocational guidance includes only the giving of information based upon a knowledge of job requirements and a careful investigation of the individual's qualifications for meeting those requirements. These qualifications, it should be noted, are defined broadly. They include not only general and specific abilities but personality characteristics as well. However, in many instances the guidance program is extended to include at least some help in vocational placement. Industrial selection is another matter which is primarily the job of the employment manager. Both in vocational guidance and in vocational placement the interest centers about the individual who is or will be seeking a job. In industrial selection the interest centers about the requirements of the job that is to be filled, and the problem is to find a suitable person to fill it. The requirements are already fixed; the only variable factor to be considered has to do with the relative qualifications of the applicants. Thus the task of industrial selection is usually less complex than is that of vocational counseling, where the range of possible vocations as well as the aptitudes and interests of the person to be advised have to be considered.

THE PLACE OF VOCATIONAL COUNSELING

Most cities make some provision for vocational counseling in the public schools. For the most part and for obvious reasons this counseling

is centered in the high schools. Inasmuch as fewer than half of the children who enter first grade remain to complete high school,¹ vocational advice for the high school student will take one of two forms: (1) educational counseling for those who expect to attend college and whose intellectual ability has been found to be at a level that justifies the expectation of college success; and (2) for those unlikely to go to college, help in choosing a career and in making preparations for it while they are still in high school.

Vocational counseling for those who do not plan to go to college will have reference chiefly to aptitude for the mechanical trades, clerical work, salesmanship, agricultural work of various kinds, minor business positions, and, for the gifted few, the various fields of the fine arts. For the girls, domestic service is added to this list.

At the college level, vocational advice will center about those fields for which college training is either requisite or highly desirable. However, as the proportion of those who attend college increases, a larger number of college graduates are entering fields in which few persons of corresponding education would have been found a generation ago. A necessary but frequently overlooked consequence of the rise in the general educational level of the population is a lessening of educational selectivity. When only a few attended college, the learned professions could absorb most of them. But as the number of college graduates outran the number of openings in the professional areas, more and more were obliged to turn to other lines of work. Thus the range of occupations concerning which the college vocational counselor must be prepared to give advice has steadily widened. At the same time the range of vocationally orientated courses that the modern college provides has increased as the need for such courses has become manifest.

In addition to high schools and colleges, a good deal of vocational advising is carried out in various psychological clinics and by psychologists in private practice. Some employment offices, both public and private, make at least an attempt to improve their services by the use of aptitude or ability tests with their clientele.

Probably the most thoroughgoing studies in the field of occupational selection that have ever been made are those carried out during World War II in connection with the choice of men for specialized military services, particularly in the Air Force. Reports of certain aspects of this work have been published by Guilford and Lacey (1947) and the

¹ According to the 1940 census report, the median amount of schooling that had been received by persons between the ages of 25-29 years was slightly less than two years of high school. Since this is based on report rather than record the figure is more likely to be high than low.

OSS Assessment Staff (1948), and in a large number of special articles in psychological journals.

QUALIFICATIONS OF THE VOCATIONAL COUNSELOR

A successful vocational counselor must not only be thoroughly grounded in the technical knowledge and skills required for his work; he must also possess certain personal qualifications that will enable him to use these skills effectively in dealing with human beings. He needs to be possessed of an extensive and fairly detailed knowledge of *job requirements*, in terms of (a) education and specialized training, (b) physical strength and freedom from special physical or sensory handicaps, (c) manual speed and dexterity, and (d) such mental and personal characteristics as ability to get on with other people, resourcefulness, stability of mood, fluency of speech, personal attractiveness, and the like. For the information of his clients he also should be well acquainted with *job opportunities*, such as usual rate of pay and opportunities for advancement, working hours and working conditions, and he should also have a reasonably detailed knowledge of the kind of work to be done. Because all these matters change with the passage of time, he cannot depend—as many do—upon facts learned in his college courses some fifteen or twenty years ago. He must keep abreast of the times. A good many aids are available for this purpose, including such periodicals as the *Occupational Index* (published by New York University, New York City) and the *Dictionary of Occupational Titles* with its 1945 *Supplement*, both of which can be obtained from the Superintendent of Documents, Washington, D.C. Together these volumes name and define briefly almost thirty-six thousand occupations. *Occupations Magazine* and *Vocational Guide* also provide a good deal of current information of interest both to the counselor and to the student.

The vocational counselor should be skilled in the techniques of interviewing, for the interview is an indispensable feature of vocational counseling. No testing program, however skillfully it may be planned and carried out, can be expected to bring out all the facts of importance for the individual case. To some extent, the interview can make up for these deficiencies.

Finally, the counselor must be well grounded in the theory and techniques of mental testing, and must have more than an elementary knowledge of statistical methods. Particularly is it important that he have developed a critical understanding of these methods as they apply to concrete situations. Far too much of the work done in this field is of

the robot type. The same may be said of the interpretations and applications of tests results. "The tests show" is a form of reasoning too often regarded as final. "The tests suggest" is a sounder way of putting it.

Among the personal characteristics of the successful counselor, the following may be regarded as of major importance. He must be genuinely interested in people. He must take a keen personal interest in those whom he advises, and feel an honest concern about their happiness and success. He should be able to meet his clients on their own ground, command their respect, and win their confidence. While he should have some of the qualities of a salesman and thus be able to present his own data in a convincing manner, he should use this approach with restraint. Decisions must always be made by the client; under no circumstances should he be coerced or persuaded to a given line of action. The counselor can help him with information about vocational opportunities and requirements. He can show him how his performances on various tests compare with each other and with those of other persons, and he can explain what they mean with reference to his chances of success in the occupational fields in which the subject may be interested. The counselor can also help, in most cases, to bolster the subject's self-confidence. Frequently he can point the way to the correction of personal habits or peculiarities that might otherwise handicap progress. But in all this his role is primarily that of providing information, not of making decisions which the subject should make for himself.

The counselor should have a good sense of order, which will enable him to plan and maintain a system of records that, if properly handled, will help him to supplement from firsthand experience the information on tests and measurements that he finds in the literature. To do this he should devise some way of keeping track of his cases. One of the most regrettable things about much of the guidance work that is carried on at the present time is that so much of its potential value is lost because of the absence of validating data. If more were known about the effectiveness of the many aptitude tests now on the market in terms of the actual counseling situation, the guidance program would have taken a long stride forward.

TESTS FOR VOCATIONAL GUIDANCE

The tests used by the vocational counselor may be classed under four general heads as follows: (1) tests of general mental and physical ability, used primarily to ascertain the subject's limitations, (2) tests of vocational interests, (3) tests of special aptitudes and abilities, and (4) measures of personality characteristics by means of paper-and-pencil

tests and perhaps by one or more of the better-standardized projective techniques such as the Rorschach or the Thematic Apperception Test. These will be supplemented by a personal history record, and one or more direct interviews designed to secure information about the subject's interests and hobbies, his major objectives, his attitude toward himself, and his opinion of his own abilities and vocational aims. Ratings by teachers and others who know the subject well are frequently added to the list.

Under the first category, a general intelligence test and a record of a recent physical examination will usually head the list. The latter may sometimes be omitted if the subject appears to be in robust health and evinces no signs of sensory handicap. For certain occupations, however, visual acuity is of such prime importance that tests should always be given. Simple tests using the Snellen charts and the Ishihara color vision tests, which can be given in the psychological laboratory, will usually detect gross abnormalities, but for more exact measures an oculist should be consulted.

The intelligence test will serve two purposes. In the high school it will serve as the most important single determinant of the wisdom of encouraging the subject to go on to college. Within broad limits, it also suggests the range of occupations within which he has a reasonably good chance of attaining success. These limits, however, are unquestionably much broader than many have supposed. The neatly organized tables that have appeared in some textbooks showing the range of IQ's suitable for various types of occupation have generally been based on academic theory rather than on statistical fact. A minimum level of intelligence below which success in a given occupation becomes highly improbable almost certainly exists, but where these limits lie has not been established in any particular case. It has been shown that students whose standing on established tests of intelligence is below the average of the general population have a poor chance of graduating from a college with high academic requirements, and since a college degree is now required for entrance into most of the professions, one may say that the equivalent of IQ 100 or a percentile rank of 50 marks the point at which success in the learned professions becomes so unlikely that it is questionable whether any person below that level, regardless of his other qualifications, should be encouraged to try to prepare for one of them. The educational hurdle alone is likely to prove an impassible barrier. Actually, standing on an intelligence test considerably beyond the level just mentioned is usually needed to bring the chances of success and failure to the 50-50 level.

This consideration brings us to an important point that bears upon

all aptitude testing. The complexities of human behavior and the intricacies of job requirements are so great that it is highly unlikely that we shall soon—if ever—be able to set absolute limits in terms of test performance for any broadly defined occupation. But when it comes to individual jobs, if these are reasonably well defined, it should be possible to establish levels of probability in terms of performance on any test or test battery designed or used as a measure of aptitude for that job. By the use of factorial analysis and multiple correlation, the accuracy of such predictions can be vastly improved, as Guilford and Lacey (1947) have shown. But in all such work the aim should be to determine the likelihood that a person making a given score on a certain test will be able to achieve reasonable success on the job in question. The results of the test should be stated in terms of the probability of success. It is because the data needed to make such calculations have so rarely been secured or reported that the need for record keeping was stressed in an earlier paragraph of this chapter. We need more facts and less propaganda; more bookkeeping and fewer unsupported claims.

Of the intelligence tests used in vocational guidance, the Wechsler-Bellevue is generally to be recommended if an individual test can be given. Not only has it the great advantage of having been standardized on a carefully selected group of adults, but its two scales (verbal and nonverbal) and the special diagnostic signs mentioned in an earlier chapter provide additional information that may be of considerable value. If only a group test can be afforded, the choice is likely to vary in accordance with the most typical characteristics of the subjects. The Terman-McNemar Group Test, the Otis, and the Wells Revision of the Army Alpha are among the most popular. The Army General Classification Test, which has recently been released for civilian use, may very possibly supersede all these, at least for the college and senior high school levels.

Among the vocational interest tests now available, the Strong Vocational Interests Tests for Men and Women (1943), of which the first edition appeared in 1927, has undoubtedly been most used, but the Kuder Preference Record (1939, revised 1942) is rapidly gaining in popularity. The Strong test is much the more elaborate of the two. The comparative simplicity of the Kuder test is one of the points that has gained favor for it. The Strong test provides separate blanks for men and for women. The men's form has keys for 36 different occupations; the women's form, keys for 24 occupations. In both forms, the occupations are generally those for which college training is either required or presupposed. The blanks for both sexes may also be scored for masculinity-femininity of interests; the men's blank for maturity and

general level of interests as well. Scores are expressed in terms of letter grades, indicating the strength of interest in each of the occupations listed. More exactly speaking, the grades indicate the extent to which the interest pattern conforms to that shown by persons successfully engaged in these occupations. A grade of A+ for "lawyer" would mean very close conformity to the typical interest pattern found for lawyers in practice; a grade of C— would mean only slight resemblance to that pattern. Most students who take this test will be graded A or A+ on one or more of the occupations; A—, B+, or B on several of the others; and B— or below on the remainder. This is in accordance with what might be expected from ordinary observation, namely, that there is not one single occupational niche for each individual but in most cases there are several into which he can be fitted without serious discomfort. As a matter of fact, Thurstone (1931), using one of the earlier forms of the Strong test, concluded on the basis of a factorial analysis that the variations in occupational interests can be accounted for in terms of four general factors: scientific interests, linguistic interests, interest in people, and interest in business. Theoretically, then, each occupational interest could be completely described in terms of a formula in which appropriate weights are assigned to each of the four factors, but whether such a procedure would save time or improve accuracy in administering and scoring the test is not certain.

The Kuder Preference Record differs from the Strong test in that it is scored for fields of interest rather than for specific occupations. The test is made up of 168 items, each of which includes short descriptions of three activities. The subject is asked to check these activities in order of preference, that is, to indicate in each case which of the three he would ordinarily like best and which least. The items cover a wide range of activities. The nine interest fields for which scores are obtained are as follows: mechanical, computational, scientific, persuasive, artistic, literary, musical, social service, and clerical. It has been found that persons in the various occupational fields show fairly marked differences in pattern of performance on this test, that college students majoring in different departments vary correspondingly, and that when the interest pattern of an individual student, as determined by this test, departs markedly from that typical of others in the department in which he is majoring, there is somewhat greater than chance probability that his classwork will not be up to the standard that might fairly be expected on the basis of his general intellectual level or that he may show signs of personal dissatisfaction and discontent.

Several other scales of occupational interest have been devised, but the two named are unquestionably the most widely used. Tiffin (1943)

has described a large number of these, together with a list of the publishers. Others are listed in the *Encyclopedia of vocational guidance* edited by Kaplan (1948). Another valuable source of information is to be found in the series of *Mental Measurement Yearbooks* (1938, 1940, 1949) edited by Oscar K. Buros and published by the Rutgers University Press. These *Yearbooks* are unique inasmuch as they include reviews by at least two competent authorities of all important tests published during the period covered, as well as data with respect to publishers, prices, etc. The reviews are critical and not always laudatory. The vocational counselor who is seeking for a frank appraisal of tests in which he may be interested will do well to consult these volumes.

TESTS OF SPECIAL VOCATIONAL ABILITIES AND APTITUDES

For obvious reasons, the task of administering a sufficiently comprehensive battery of separate tests capable of measuring a subject's ability along all possible lines would be too great to be seriously considered. Several ways of solving the difficulty have been employed. In advising high school students who are undecided as to their vocational choice, the method most commonly used is to begin by ascertaining something about their intellectual abilities which will serve as a guide to the general occupational level into which they are most likely to fit. If a student's intelligence is reasonably high, one of the tests for determining his general pattern of occupational interests is given. An interview with the student, and, if possible, with his parents, then follows, on the basis of which tentative plans having to do with his immediate educational preparations can be made.² More specific decisions, in these cases, can be delayed until college entrance or even until after one or two years of college training have been completed.³

For students who do not plan to attend college, vocational choices are called for at an earlier age and must be made in more specific terms. The Intermediate Form of the Kuder Preference Record (1944) is of some help here, but the results should be checked with the student's

² Such, for example, as whether his high school courses shall emphasize the sciences or the arts, whether his language courses shall be mainly Latin and possibly Greek, or if modern foreign languages should be substituted. The choice of a college will also depend, to some extent, upon the student's field of major interest.

³ In many colleges and universities, final decision as to choice of a major is not made until the end of the sophomore year, although a large proportion of the students will have made their choice at a much earlier period. In these schools, the curriculum for the first two years is not highly differentiated but is made up largely of courses thought desirable for all.

expressed interests and with the evidence provided by his school record. The question that needs to be decided fairly early in his high school career is whether his program shall stress business and commercial courses, shopwork leading to one or another of the mechanical trades, general science, with a possible view to farming or to certain types of technical positions for which college training is not essential, or, for girls, to domestic service work, beauty culture, and the like. The interest tests should be followed by tests which will provide at least a partial check on the results. A comparison of the score on the Minnesota Clerical Abilities Test (Andrew, *et al.*, 1941), with the combined score on the Minnesota Paper Form Board Test (Revision by Likert and Quasha, 1934), and the O'Rourke Mechanical Aptitude Test, Junior Grade (1937) may provide a reasonably satisfactory check on the separation between the two broad fields of clerical aptitude on the one hand and mechanical aptitude on the other. If the two lines of approach (interest as shown by the occupational profile on the Kuder, and aptitude as shown by the comparison of the two kinds of tested abilities) are in conformity with each other, a choice of high school curriculum will be suggested for a large number of the cases. If they disagree or if scores in both areas are low, it is likely that some other vocational field is indicated. Again the preference blank may be consulted for possible clues. Artistic and musical interests as there indicated can best be checked (as far as aptitude is concerned) by performance, since no really good tests for predicting success in these fields are available. Most school systems, however, provide a sufficient amount of art training to make the student's performance along these lines of some value for judging his ability. Many schools also make some provision for training the more able students in instrumental as well as in vocal music, a good many of whom will have had additional private instruction in music. But vocationally speaking, neither of these areas will provide gainful employment for more than a small percentage of cases.

Kuder's "persuasive" and "social service" areas have no very exact counterparts among the aptitude tests. Some of the more dependable of the personality tests, particularly those designed to identify the socially dominant as opposed to the submissive type of individual such as the Guilford-Martin Inventory of Factors G A M I N (Martin, 1945), should be related to the former.⁴ Generally speaking, however, the personality tests thus far developed are likely to be of more value in helping to

⁴ The five factors of this inventory are defined as follows: G—general pressure for overt activity; A—ascendancy in social situations (as opposed to submissiveness); M—masculinity of interests and attitudes; I—absence of inferiority feelings, that is, self-confidence; N—lack of nervous tension.

identify persons who are unsuited for certain kinds of work than for use on the more positive side of vocational guidance. Good social qualities, for example, are an asset almost everywhere, but there are certain types of work for which such qualities are so essential that one who is lacking in them is doomed to failure at the outset. In others they may count for less. A shy and timid policeman with a submissive rather than a dominant personality is not likely to be of much value to the force though he might be a very satisfactory bookkeeper. A forest ranger may be pretty much of an introvert; a stage comedian, rather an extreme extrovert, and both may be happy and successful in spite of personality characteristics that would be a handicap in many types of positions. In view of the many questionable features of most of the personality inventories, the vocational counselor should not depend too much on them. However, if interpretations are made with restraint and are checked from other sources, such as reports of associates and by personal interview, two valuable purposes may be served by these inventories. First, they may be used as screening devices for the selection of students who are in need of help with personal problems that should be cleared up as far as possible before vocational advice can be wisely given. In the second place, when verified from other sources, the information thus obtained becomes a valuable supplement to tests of interest and ability, since it may serve to modify or to verify the conclusions thereby drawn; or, when it comes to the choice of specific vocations, the personality variable may be an important factor in deciding among those within the general field to which other measures point.

PART IV

Applications

Testing in Schools and Colleges

THE MAJOR OBJECTIVE

Mental measurement in schools and colleges is designed to further the end toward which all education is directed—that of furnishing the best possible opportunity for the growth and development of the individual student.

The agricultural worker who wishes to provide the best possible conditions for the growth of his plants needs several kinds of information. First of all he must know his plants and must ascertain their individual requirements of soil, moisture, temperature, sunlight, and other conditions. He must be able to measure these conditions in order to know whether or not the proper degree of each has been obtained. Finally he must be able to test the results of the methods he uses in order to determine whether or not the assumptions he has made were correct.

The education of children and youth involves much the same problems. If progress in the art of teaching is to be enhanced at anything like the rate that has been achieved in the biological and agricultural arts or in the field of engineering, a firm grounding of educational method in scientific fact is essential. This means that we must know the human subjects with whom we have to deal. Although, unlike the plant and animal breeders, we cannot hope to improve the basic qualities of our stock, we can learn to classify it. We can find out what kind of educational opportunities best fit the needs of each class. We can learn to define these opportunities more exactly and thus be able to apportion them more appropriately. We can study the results of our attempts and thereby direct our later efforts more wisely.

For each of these aspects of the educational process, measurement is required. Not only must we be able to describe the students with whom we work in more exact and detailed terms than unaided human observation makes possible; we must also be able to analyze and describe the environmental conditions to which they have been subjected in the past and the new conditions which we intentionally impose in the

course of our educational program. The effect of these conditions can then be ascertained in terms of changes in the students when appropriate allowance is made for individual differences among them and for the presence or absence of environmental factors that may have given rise to these differences.

THE APPLICATIONS OF MEASUREMENT TO EDUCATIONAL PURPOSES

It is unfortunately true that much of the testing done in schools and colleges is carried on without a very clearly defined purpose or with a purpose of such limited scope that only a few students are clearly affected and the findings for the majority are wasted. Another unfortunate practice arises from the erroneous assumption that teachers are sufficiently informed about the meaning and use of tests to enable them to profit, without further instruction, from a knowledge of the results of those given to their students. The findings of school surveys are sent to the office of the school principal where teachers may have access to these results of mental and educational tests. Highly regrettable consequences often follow. Children may be praised or blamed for their standing on intelligence tests, or the results of these tests may be reported to their parents. The low standing of particular children may be taken as an excuse for ceasing to make further effort with them—they are, it is said, “too stupid to learn in any case.” These and similar practices, all too common even among the comparatively enlightened school personnel of today, make it imperative that if the testing schedule is to fulfill the purposes it is potentially capable of serving, *teachers must be given more adequate information and training in the meaning and uses of test results.*

The courses given in most schools of education are either too specific or too general to meet this purpose. They are directed toward the prospective psychometrist or research statistician or they accomplish little beyond promoting a certain glibness in the use of technical expressions which are but imperfectly understood by those who use them. It is not necessary for the average classroom teacher to become highly skilled in the more complex aspects of test administration or clinical research. But as long as the educational process is carried out through the instrumentality of teachers, they must be taught to make the best possible use of the information provided by scientific workers in those fields from which education derives the material for its theories and methods. Here the area of tests and measurements holds an important place.

In the application of measurement to the service of education a

number of subsidiary objectives may be noted. Among these are the following:

1. To promote the general adjustment of the individual student through better understanding of his abilities, aptitudes, capacities, and needs. By *abilities* we mean the knowledge and skill which he has already acquired along various specified lines. *Aptitude* refers to the readiness with which he increases his knowledge and improves his skill under specified conditions of opportunity and training. *Capacity* is indicated by measurement, or, more often, by an estimate based upon measurement, of his physical or mental limitations for further attainment under existing conditions or after these conditions have been altered in some specified way. *Needs* refer to those aspects of his environment which are deficient in respect to one or more of the attributes requisite for his optimal development. From the combined results of these four factors, interacting with previous experience and present circumstances, the *conduct* of the individual, with its associated features of interests and desires, feelings and emotions, habits and attitudes, is derived.

2. To aid in the identification of extreme deviates for whom special methods of instruction may be needed.

3. To aid grade placement or the classification of students into ability groupings, or to serve as a partial standard for college admission.

4. To provide data for use in educational and vocational guidance.

5. To provide objective tests of the relative effectiveness of different ways of presenting school subject matter or different plans of classroom organization.

Many of our current educational practices are based upon unproven assumptions, traditional beliefs, and sheer laziness. Education, we have said, is an art, not a science, and we have straightway proceeded to demonstrate that at least half of this statement is true! But no practical art can progress rapidly unless it is grounded in a thorough system of facts and principles which can be applied to the actual situations that arise. Intuition is not enough. Personal experience, while helpful, is neither sufficiently extensive nor sufficiently well organized and free from bias to be a safe guide. The farmer who plants his crops in whatever location chances to strike his fancy and who fertilizes and cultivates all of them in the same manner, regardless of their kind, is unlikely to have a good harvest. The one who makes use of the results of agricultural experiments on plant needs, tests his soil, and chooses the varieties of plants that will do well under the conditions he is able to provide has a much better chance of success.

Educational experimentation is not a fad. It is a field of scientific inquiry that reaches to the very roots of our civilization, for on the

character of our youth depends the future of the race. Its history is not long; until practical methods of mental measurement had been developed, the results of such experiments could not be expressed in quantitative terms. The likelihood that similar results would be obtained when the experiment was repeated could be tested only by the laborious and costly methods of empirical trial. But with the development of feasible methods by which changes in behavior as well as growth in physical structure could be subjected to quantitative appraisal, together with improved statistical devices for analyzing and synthesizing the measures thus obtained, experimental education ceased to be merely a high-flown term for use by erudite speakers at teachers' conventions. It became an established fact.

The use of tests and measurements in modern education is far too broad a subject to be handled in a single chapter. The *Review of Educational Research* publishes triennial summaries of the investigations carried out in a number of important areas. Each of these summaries is followed by a fairly comprehensive bibliography. Each of the yearbooks of the National Society for the Study of Education is devoted to a single major topic of educational import. Both discussions of current theories and summaries of the experimental literature in the field are usually included, in addition to original reports of new investigations. As a rule, two volumes, each on a separate topic, are published annually. Some of the more recent titles of these volumes are as follows: *Juvenile delinquency and the schools* (Forty-seventh Yearbook, Part I, 1948); *Reading in the high school and college* (Forty-seventh Yearbook, Part II, 1948); *Science education in American schools* (Forty-sixth Yearbook, Part I, 1947); *Early childhood education* (Forty-sixth Yearbook, Part II, 1947); *The measurement of understanding* (Forty-fifth Yearbook, Part I, 1945). Ross (1944) has prepared an outline and description of ways in which tests and measurements are used in public schools, and a committee headed by J. G. Darley has published a general discussion of the use of tests in colleges. This list might be extended almost indefinitely, but the examples given will provide an idea of some of the many and varied uses of tests and measurements in modern education, as well as of the broadened concept of the nature of the educational process, in the development of which these tests have played a considerable part.

Testing in Clinical Practice

VARIABLE FACTORS AFFECTING THE SELECTION AND USE OF TESTS

As has been noted in previous chapters, psychologists differ in their points of view with respect to many aspects of mental testing. Some are ardent supporters of the holistic approach to the study of human nature and conduct. They believe that human behavior is the end product of a well-nigh infinite number of antecedent factors acting in organized unison. These factors cannot be measured directly; they can only be inferred from the behavior to which they lead. But inasmuch as the behavior depends primarily upon the organization, the relationship of the various factors to each other and to the goal toward which the behavior tends, it can only be studied *in toto*, for a change in the organization of the parts means a corresponding change in the behavior, even though the factors themselves are unaltered.

The testing procedures used by psychologists who accept the point of view mentioned above differ in many respects from those favored by adherents to the theory of measurable traits. The former tend to favor projective methods or other devices from which descriptive characterizations of the subjects are obtained. Person-to-person measures, such as Moreno's sociometric index or other measures designed to test the total impression that an individual makes upon his associates, also meet with their approval. In general they are more concerned with attempts to predict conduct than with the measurement of ability, although many depart sufficiently from their own theoretical position to make use of standard intelligence tests. In general, however, clinical organizations favoring the holistic approach are less concerned with statistical than with observational verification of their conclusions. They lean heavily upon case reports for support of their procedures.

Differing from this group in many respects are those who approach the study of human beings from an analytic point of view, and whose

test of a measuring device is made in terms of group statistics based upon methods in which the variable factor of human judgment is eliminated as far as possible. Upon the assumption that human conduct depends upon, and is consistent with, a basic underlying structure of abilities, habits, attitudes, and motives which may change with time but remain comparatively constant over short periods, an attempt is made to develop instruments of precision for measuring these attributes. As evidences of the accuracy of these measures, coefficients of self-correlation and correlations with other criteria are cited. Case reports are used chiefly as illustrative material.

The character of the testing program in any clinic will thus depend to a considerable extent upon the theoretical position of those who are in charge of it. It is true that in many, perhaps in the majority of cases, neither of the two views just outlined will be held to the complete exclusion of the other, and the methods used will include some that are appropriate to each. As a rule, however, a definite leaning toward one or the other standpoint will be apparent both in the selection of tests to be given and in the interpretations that are made of the results.

Not only the theoretical position of its director but the clientele which the clinic is designed to serve and the agency responsible for its financial support will also have a bearing upon the aims of the testing program and the kind of tests included within it. Among the organizations which deal only with adults are some, in many cases connected with mental hospitals, which are mainly concerned with psychiatric diagnosis and therapy for patients with pronounced mental or emotional disturbances. The tests used will be mainly directed toward that end, although they may include some measures of general ability and vocational aptitude as a guide to occupational therapy and later assistance in job placement. Others, particularly those which are or attempt to be self-supporting (including private psychological practitioners), usually emphasize vocational guidance as their main service to adult clients, although other types of advice, dealing with such personal problems as marital difficulties, personal worries, and social problems, are also given. Tests designed to aid in uncovering the source of such difficulties include word-association procedures, the Rorschach, and the TAT, as well as many of the paper-and-pencil inventories, the Burgess and Cottrell or the Terman tests of marital happiness, and others selected in terms of what is known about the particular case. Interviews and life histories are also looked to for clues. During recent years, much interest has been displayed in the nondirective interviewing procedures for the development of which Carl Rogers (1942) has been chiefly responsible. The establishment of clinical services to aid in the rehabilitation of the

handicapped has been an important feature of the Veterans Administration program. For this work the psychologist requires a thorough grounding in the use of tests designed for the blind and deaf and for those suffering from other types of physical handicap. He must be on the alert to detect signs from which the adequacy of the test performance of a given person on a particular occasion can be inferred. He should likewise be well informed about vocational openings for different classes of handicapped persons.

Clinics for children and adolescents may be divided roughly into the following classes: (1) those supported by the public schools; (2) those supported by a state in connection with its department of institutions and agencies;¹ (3) clinical or research divisions within the larger institutions, such as schools for the feeble-minded or reformatory schools for young delinquents; (4) child-guidance clinics for the diagnosis and (in some cases) treatment of various forms of emotional and behavioral problems of children and adolescents. In these as well as in the organizations dealing chiefly with adults, the testing program will vary not only with the interests and points of view of the clinic staff but also, and of necessity, with the type of clientele served.

PSYCHOLOGICAL TESTING IN PSYCHIATRIC CLINICS FOR ADULTS

The role of the psychologist in the psychiatric clinic in the past has too often been looked upon as that of a technician or a laboratory assistant rather than that of a scientist, expert in his own field. This view has been responsible for the fact that many clinics have employed as "psychologists" persons of mediocre ability whose training did not exceed one or two years of graduate work that included a course or so in mental testing. In many cases, these persons could not meet any reasonable standards set for psychometrists, let alone clinical psychologists. As long as no formal professional standards for the title had been established, there was little that could be done to remedy this situation, but with the new requirements for certification set up by the American Board of Examiners in Psychology, some improvement, at least, should be brought about.

The work of the psychologist in a psychiatric clinic² supplements

¹ The name given to this department varies from one state to another, but most states have some division that recognizably corresponds to it.

² Assuming, of course, that he is a psychologist, with the necessary training and experience to merit the title. This ordinarily means from three to five years of graduate work leading to the Ph.D., together with a sufficient amount of practical experience for orientation in the work to be done.

that of the psychiatrist (1) by adding to the patient's case report the records of his performance on a series of objective tests, (2) by making careful observations of the patient's behavior during the test situation, which is thus regarded in part as a standardized experimental setup in which the reactions of different subjects under uniform external conditions may be noted for comparison, (3) by conducting special experiments such as word-association tests with or without such mechanical aids as the galvanometer or the various devices for recording changes in muscular tension to aid in uncovering factors which underlie the patient's emotional disturbances, (4) by summarizing this material and organizing it in such a way that significant points may readily be brought out for discussion at the staff conference, and (5) by participating in these conferences. He will also, as a rule, be largely responsible for securing information about the occupational experiences, aptitudes, and interests of the patients which can be used as a guide to occupational therapy while they are still in the hospital or as a help in job placement for those who are discharged. Finally, in view of the great need for additional data as to the usefulness of the methods he employs, particularly those involving special diagnostic tests or behavioral signs of diagnostic significance, every psychologist who is carrying on work of this kind should regard it as his professional duty to make as much of his experience as possible available to his colleagues by means of published reports. These studies may be carried on independently or in cooperation with the psychiatrist with whom he is associated.

PSYCHOLOGICAL TESTING IN BEHAVIOR CLINICS FOR CHILDREN

Inasmuch as the intellectual level of the child is one of the main determining factors in laying out a corrective program for him, the administration of one or more tests of general mental ability is usually regarded as the first step in the examination schedule for each child. Because the Stanford-Binet covers the complete age range likely to be found among the clientele of a child-guidance clinic and is generally regarded as the most dependable of the tests of its kind, its use has become standard practice among most clinicians. It is desirable, too, to have at least one basic test that is given to all subjects, regardless of their age or the nature of their difficulty, which can serve as a general reference point in comparative studies of different individuals or contrasted groups.

As a check on the predominantly verbal Stanford-Binet, a nonverbal test of some kind is frequently given, especially if the child in question

is deficient in the use of English, has a speech defect, or has appeared to be emotionally inhibited in regard to his verbal responses when taking the Stanford-Binet. The test used will depend upon the level of maturity of the subject. Generally speaking, the Merrill-Palmer will be chosen for children of preschool age, the Arthur for elementary school children, and the performance scale of the Wechsler-Bellevue for high school children. However, it is always necessary to keep in mind that these tests are less dependable³ for predictive purposes than is the Stanford-Binet. Unfortunately common among many clinicians is the practice of taking it for granted that, if the subject does better on the non-verbal test than he did on the Stanford-Binet, sufficient evidence is thereby afforded that the latter did not give a truly representative picture of his ability. But such an interpretation is not warranted unless it can be shown that the divergence between the two is greater than can reasonably be accounted for on the basis of the experimental errors of measurement or unless enough supporting evidence can be adduced to justify the assumption that the score on the nonverbal measure is in this case the more dependable of the two. The mere fact that it is higher does not in itself warrant such an assumption.

Another practice of the poorly trained clinician should be noted here. Upon the assumption that the standard instructions for administering certain items of the Stanford-Binet unfairly handicap children from non-English-speaking homes or others whose experiences or personal characteristics set them off from the generality, these persons attempt to make up for the difficulty by changing the standard procedure to one which they deem more suitable for these exceptional cases. As a rule, no record of these changes is made. The result of the altered test is taken at its face value.

Now the examiner's assumption that, because of the special features of the particular case, the test in question may not provide a truthful picture of the subject's ability may be wholly correct. If this is true, two alternatives are open to him: he may substitute some other test, or he may give the first test *according to the standard method and write in a conspicuous place on the front of the blank his estimate of what the subject's ability would have been rated by means of a wholly suitable test*. If he has sufficient faith in his own judgment to warrant a change in the manner of giving the test, he should be willing to go on record to that effect. But if he uses the form of the test to bolster an opinion that he realizes has not sufficient grounding to stand by itself, there is no

³ With the possible exception of the Wechsler-Bellevue, about which less is known for long-time prediction because of the comparatively short time since its publication.

way by which the two can be disentangled later on. The observations and judgments of an experienced clinician are valuable aids to test interpretation and as such should always be recorded, but they should never be allowed to confuse or obscure the test results.

In addition to the tests of general mental ability, others will be added to the schedule as indicated by the nature of the child's difficulty. Because the school occupies so large a place in the child's life, educational tests will be desirable in most cases. These should include as a minimum (1) a test of speed and comprehension of silent reading material, (2) an arithmetic test, (3) a spelling test, preferably in the form of dictated sentences which will provide a rough indication of handwriting skill as well. Special diagnostic tests designed to analyze the nature of reading deficiency or specialized difficulty in other subjects should be given when indicated.

Cases referred because of emotional or social difficulties, or as a result of delinquent behavior or other types of conflict with authority, require study from two angles, in addition to whatever clues may be given by the intelligence tests or the tests of educational achievement.⁴ First, it is necessary to get as unbiased and detailed a picture as can be had of the child's home and school background, and of any major incidents in his experience that may have been disturbing factors in his emotional and social life. This information will be obtained from a number of sources. First there will be the social case history, usually obtained by a social worker connected with the clinic. The psychologist may also wish to interview the parents with respect to some of the points brought out in this history. One or more of the inventories on parent-child relationships, such as that by Stogdill (1934), which deals particularly with parents' attitudes toward the amount of personal freedom which should be permitted children, may be well worth using.⁵ The

⁴ In a good many cases, a marked discrepancy will be found to exist between the grade in which a child is located in school and that which is indicated by his level of intelligence and his educational achievement. Such discrepancies almost inevitably accentuate a child's difficulties, even though they may be contributing factors rather than major causes.

⁵ An unpublished study by Katherine Miles (Ph.D. thesis, University of Minnesota) describes a number of interesting approaches to the study of parents' attitudes toward their children. One of these, called the "Choose Your Child" test, is especially original. A series of descriptive sketches, each of which depicts the characteristics of a hypothetical child, was prepared, and the parents were asked to arrange them in order of preference. The characteristics dealt with such factors as submissiveness vs. independence of thought and action, fondness for companions of their own age as compared to preference for the society of their parents, adventurousness vs. timidity, etc. A second questionnaire dealt with the opinions of parents with reference to a wide variety of items on child training and management. These were filled out independently by the fathers and mothers of a group of children chosen on the basis of judg-

Fels scale for rating parent behavior (Champney, 1941) filled out by the social worker is another possible approach. The Leahy scale for rating urban home environment (1936) provides fairly objective data on the material and cultural aspects of the home, together with some information on the social activities of the parents.

The second important angle to be considered is the child's attitude toward his world and toward himself as a part of it. Does he feel secure and happy in his relations with his home and with his school and playmates? Does he consider that he is fairly treated? Does he feel resentful and rebellious toward certain persons or organizations, and if so, what is the basis for this feeling? It is necessary always to remember that the objective facts of a particular situation or episode as seen by others are of less significance here than is the child's perception of those facts and his reactions to them. The Bell School Inventory (1939) and Stott's Home Adjustment Inventory for use with adolescents (1941) are likely to throw some light on the child's view of the two institutions in which the major part of his life is normally spent. Used as a guide to interviewing, the specific items of these inventories will often point the way to much more valuable data than are given by the quantitative scores. With proper recognition of their hazards and limitations, experiments using some of the projective techniques described in a previous chapter may also provide useful clues.

The basic program for children sent to a behavior clinic will include a physical examination or a report from the family physician as to the child's general health, together with any special medical examinations or tests of sensory acuity that seem called for, a social case history, and one or more interviews with the psychiatrist if the clinic is under psychiatric direction. These materials will be available to the psychologist, who will use them as a partial guide in organizing the individual testing schedule. The organization of case reports in such a way as to bring all the salient facts into clear relief is also likely to be a part of the psychologist's work, since the chances are that there will be no other person in the organization who is skilled in preparing such records. Considerable time and thought may well go into the planning of face sheets and summaries, for they will become important tools for the research that

ments by their high school classmates to represent wide differences in social behavior, especially those aspects having to do with leadership. Marked differences in parental attitudes were found. In general these differences paralleled the behavior of the children. Children with strongly developed social abilities and who tended to be the leaders of their groups usually had parents who respected their personal rights and encouraged independence. In contrast, the children of dominating parents rarely showed traits of leadership and were frequently withdrawn and solitary in their behavior.

should be regarded as one of the most important parts of the services rendered by the clinic.

Repeated investigations have shown that the case reports of outstanding success in the treatment of children's behavior problems that so frequently appear in the literature are by no means representative of what is typically achieved, even in those clinics with exceptionally well-trained staffs and relatively ample facilities for therapeutic treatment (Witmer, 1940). The fact that parents are often obdurate in their refusal to accept suggestions or that home conditions are such as to defeat the most earnest efforts to improve the child's behavior should be looked upon as a challenge rather than as a barrier. We need to know how to overcome conditions that have stood in the way of success. To this end, the following facts need to be marshaled for comparison: (1) the characteristics and behavior of the child at the time the study is begun; (2) the major objective facts concerning his environment and personal history; (3) the child's attitude with respect to those facts, how he regards the world in relation to himself; (4) the kind of therapy used; (5) a detailed progress record which includes information concerning changes in the environment, especially in respect to modifications in the attitude of parents and teachers toward the child and in his relations with his companions. In the nature of the case, much of this material will be descriptive rather than quantitative, but repetitions of some of the tests used earlier will provide a certain amount of validating material.

Witmer (1946) calls attention to the changes in clinical procedures and points of view that have occurred during recent years. Because of the great difficulties involved in the former attempts to help the child by modifying his environment, and the consequently large numbers of cases for whom no apparent improvement was brought about⁶ in spite of earnest effort on the part of the clinic staff, a different approach is now being tried in many clinics. Instead of centering efforts on attempts to help the child by reducing the external stresses to which he is subjected, that is, by modifying his environment, the psychiatrist attempts to build up an intimate relationship between himself and the patient and to use this relationship for therapeutic ends. The aim is not to simplify the child's problems from without but to change his attitude toward them and

⁶ In her 1940 report, Witmer cites the following figures for certain clinics in northern New Jersey with the comment that these are somewhat more favorable than those found in other clinics. Within a random sample of 200 cases whose condition was studied either at the close of treatment or for whom treatment was still in progress after a period of not less than eighteen months after admission, 12 per cent were adjusting satisfactorily, an additional 27 per cent showed some improvement, while 24 per cent were unimproved. The remainder had either been lost track of or, in a small number of cases, been sent to institutions.

toward himself in such a way that they will no longer oppress him unbearably. The ten cases used by Witmer to illustrate how this change is brought about were purposely selected because the method appeared to be successful with them. As she is careful to point out, they cannot be regarded as typical in any sense of the word. Whether or not the newer methods produce better results, on the average, than did the older ones is uncertain. Here, too, research is needed.

THE ROLE OF THE PSYCHOLOGIST IN REHABILITATION WORK WITH ADULTS AND OLDER ADOLESCENTS

The close of World War II left us with a tremendous increase in the number of physically handicapped and emotionally exhausted men who must be helped to live as normal a life as is possible for them. The challenge that this situation offers to the clinical psychologist is inexpressibly great.

Rehabilitation of the handicapped person involves much more than attention to his physical ailments, important as these unquestionably are. It means help in learning to accept his condition. The handicapped person, no less than the normal one, must look upon the future with hope and confidence. He must adopt a friendly and tolerant attitude toward the world, and feel himself to be an essential part of it, not merely a burden upon it. He must find a place which is satisfying to himself and in which he knows himself to be a source of satisfaction to others.

That this is not an easy task is self-evident; it is an eloquent testimony to the courage of the human race that so many are accomplishing it with little help. Others less fortunately endowed or more unfortunately situated are unable to cope with their difficulties unaided. It is here that the clinical psychologist can be of help.

Under the Veterans Administration a large number of clinics and vocational counseling centers have been set up. Many of these are located in the veterans' hospitals; others are independent units. The staff usually includes both psychiatrists and psychologists as well as other medical officers with many fields of specialization. The services performed by the psychologists in these clinics are many and varied. They have to do with the personal problems of the patient as well as with his vocational guidance and placement. The large number of war marriages entered into with but slight mutual acquaintance and little consideration of the future have given rise to many difficulties of marital adjustment, often complicated by the arrival of unplanned-for children. War injuries and

war strain left their marks on the personality and emotional make-up of many men whose bodies suffered little or no lasting damage.

In the work with these persons, mental testing has an important role. In general, tests are used for three purposes: (1) diagnosis and classification with special emphasis on vocational aptitudes and the identification of areas for which special training or psychiatric help is needed, (2) prognosis, and (3) the measurement of progress under treatment. Cooperative research in which the medical and psychological aspects of the work are brought into organized coordination can be of much value here.

MENTAL TESTING OF PHYSICALLY HANDICAPPED CHILDREN

Pintner and others (1941) have given an excellent survey of the methods and results of testing the physically handicapped which need not be repeated here. The bibliographies at the end of each chapter will be very helpful for those especially interested in this field. In this brief section we shall mention only one or two points that have perhaps been insufficiently stressed by these authors.

The physically handicapped child, even more than the normal child of corresponding age, is dependent upon his parents for mental and emotional nourishment. Their attitude toward him may be stimulating and encouraging, or it may lead to increased dependence and self-pity. They may regard the child's affliction as something that primarily affects themselves ("Why should *I* have this burden laid upon me?") or they may, with more or less reason, blame themselves for his condition ("Why did I not guard him more carefully?") and try to expiate their feeling of guilt by lavishing unnecessary time and attention on the one whom they believe they have wronged.

The study of parent-child relationships is of particular importance in these cases. Mental tests to determine the child's potentialities as an aid to directing his education and training should therefore be supplemented by tests designed to study the parents' attitudes toward the child and in many instances to determine as well as possible the effect that his condition has had upon the family life and the relations between the parents. Tests for this purpose are still pretty much in the experimental stage, and the alert psychologist may perhaps wish to try out some of his own devising. However, some of the inventories previously mentioned, particularly when used as a partial guide to subsequent interviews, are likely to be informative, while such tests as the TAT or the Rorschach may also yield suggestions that are helpful.

SPECIAL PROBLEMS OF THE CLINICAL PSYCHOLOGIST

One of the major difficulties that the clinical psychologist faces arises from the well-nigh infinite variety of the problems he is called upon to solve. The research worker in the psychological laboratory chooses a single problem. He devises a plan of attack and selects his subjects in accordance with that plan. He interprets his results with reference to his original problem.

The clinical situation is very different. The psychologist cannot choose his problem; it is thrust upon him. He cannot, as a rule, devise a carefully set up experimental situation and spend long weeks or months in working out his answer. The pressure of new cases with different problems is too great.

All this has tended to deter progress in clinical procedure by preventing the adequate testing of hypotheses and the validation of methods through sound experiment. The greater number of the reported studies in the field are deficient in suitable controls. Few long-time studies covering the same subjects have been made. The fact that very few clinics make provision for following their subjects after the case has been officially "closed" inevitably makes for a somewhat superficial view of the effectiveness of the treatment given.

Lack of funds and consequent shortage of adequately trained personnel, together with the constant pressure of new cases clamoring for attention, are, of course, mainly responsible for this situation. Much, however, could be accomplished if a group of clinical psychologists in different organizations handling much the same type of clientele were to arrange a uniform system of records for their cases and agree on a series of problems of mutual interest to be attacked jointly. Such problems might involve the usefulness of certain tests for specified purposes, or the testing of hypotheses with respect to the association between measured characteristics of the subjects tested and special features of the environment.

The clinical psychologist learns by experience, it is true. But experiments carried out in the psychological laboratory have convincingly demonstrated that incidental learning is a slow and laborious process, in the course of which chance associations are likely to be confused with fundamental principles, and beliefs may be formed on the basis of wishful thinking rather than on demonstrated fact. That many clinical practices have been developed on no sounder basis than this is unquestionably true. That most clinical psychologists fail to gain much that is potentially possible out of their varying experiences is also undeniable. And

since the future of clinical psychology hinges upon the psychologists who practice it, the pivotal question is: How can the experience of these persons be organized in such manner as to provide new and sounder information by which to guide the work of those who will come after them?

A complete answer to this question cannot be given at the present time, but certain aspects of it are clear. Experience can be transmitted to others in two ways: by demonstration and the spoken word to a few, and by written records to many. More carefully planned systems of record keeping therefore appear to be one of the methods whereby experience can be made available for others. But these records will not serve a useful purpose except as they are brought together and used for the testing of specific hypotheses. Clinically oriented research of this kind is much needed.

The Use of Tests in Industry

PURPOSES FOR WHICH TESTS ARE USED

As originally used in industrial plants, tests were regarded chiefly as devices for reducing employee turnover through better selection of candidates for particular kinds of jobs. When many applicants for each job are available, industrial selection is still likely to be one of the main uses to which tests are put. In times of labor shortage, however, when the number of available jobs exceeds the number of men available to fill them, the emphasis is shifted from the requirements of the job to the abilities of the men. Although the man-job relationship is still the field of inquiry, the immediate problem is no longer merely that of finding the best man for a particular job, but is likely to be that of finding which one of a number of jobs that need to be filled will be the most suitable place for a particular man. Even in normal times, the efficient plant superintendent or personnel manager is likely to find that misplacement on a job is a frequent source of discontent in employees, and that a more careful fitting of the man to the job results in a more smoothly running organization. Although tests are by no means infallible guides, either to employee selection or to industrial placement of employees already chosen, experience has shown that when wisely used they can be of material help for both purposes.

The personnel manager, whose task it is to help to reduce plant friction through the satisfaction of individual complaints and to improve plant morale by straightening out misunderstandings between employees and supervisors, will also find that tests rank among the most important tools of his profession. Not only will their use be an important means of helping to explain individual cases of discontent arising from job misplacement but they may enable him to advise the plant manager with respect to positions for which the men in question are better suited. The personnel manager can help to avoid unnecessary wastage arising from the employment of competent men on mediocre jobs. He can help to locate such focal points of trouble as the chronic malcontent whom no

ordinary conditions will satisfy,¹ the scandalmonger who breeds trouble among his mates as the sparks fly upward, the paranoid who sees evil intent behind the most ordinary actions of those about him, and other maladjusted workers whose private difficulties give rise to public problems. Such persons may be handled in one of three ways. For the milder cases whose problems are not very deep-seated and in whom the behavior manifestations are fairly recent, direct therapeutic work with the individual himself may be feasible. Sometimes it may be possible to transfer the troublemaker to some other department or position where his opportunities for creating difficulty are fewer. Dismissal may be necessary when neither plan is feasible.

Particularly in times of labor shortage, when it often becomes impossible to secure a sufficient number of experienced workers to fill positions as they become vacant, arrangements must be made for the training of new workers within the plant itself. From the standpoint of industrial efficiency four problems are involved here. Since no previous work record is available for these candidates, few of whom will ever have been employed before, the problem of general promise must be considered. There is likewise the question of specific aptitudes. For what particular kind of job shall each be trained? And since neither human judgment nor available tests can provide an infallible answer to either of these questions, the progress of each candidate must be measured from time to time in order that the soundness of the original decisions may be checked and errors corrected without unnecessary waste of time. Finally some method is needed to determine which members of the plant personnel are best qualified to provide the necessary training. Not all the men who are themselves expert in a given line of work can impart their knowledge and skill to others, nor are all equally capable of inspiring these young people with a feeling of loyalty to their jobs, their employers, and their fellow workers. This is an aspect of training too often overlooked.

Finally, tests can be of much value in helping to determine the physical conditions that make for the workers' efficiency and happiness, and the type of working program that will make the most effective use of their working time. Many of the efficiency studies reported in the literature have disregarded an important feature of overhead costs when com-

¹ Such a person is usually suffering from some quite deep-seated personality difficulty. He may be in need of psychiatric treatment which the personnel worker is unable to give, either from lack of the special training needed or because of lack of time and facilities. A few of the larger industrial concerns provide psychiatric services for their employees but as a rule, unless the condition proves to be very mild, all that the personnel worker can do is to advise treatment by a psychiatrist outside the plant or to effect such changes in working conditions as may serve, to some extent, to alleviate his complaints.

paring output under different conditions of work: they have failed to take account of labor turnover. The advantages imputed to a given method may quickly vanish if it is found that a procedure which yields increased output per man-hour is at the same time so distasteful to the men that they start looking for new jobs. The effect of a given set of working conditions upon the attitudes of the men toward their jobs becomes one of the most important variables to be taken into account in any type of industrial engineering. Attitude tests especially constructed to meet the needs of the immediate situation, as well as some of the best of the standard personality inventories, may be helpful here if used with discretion.

TESTS FOR INDUSTRIAL SELECTION AND CLASSIFICATION

The tests chiefly used by the industrial psychologist will fall under the following broad classes: (1) tests of general mental ability; (2) tests to determine relative ability for the four main classes of industrial work: administrative work (including foremen and supervisors, office managers, and the like),² clerical work, selling, and shopwork, including not only machine operating but also assembling, inspecting, and packing; (3) tests to determine aptitude for specific types of work under each of the above general classes; (4) achievement tests for use in measuring progress in the acquisition of a particular type of skill; (5) tests of personal-social characteristics, including measures of emotional imbalance and neurotic tendencies.

Tests of general mental ability will be used for two purposes: to eliminate candidates unsuitable for employment in any capacity because of their low level of general understanding, and as a supplement to tests of special ability in determining not only the kind but the general level of employment for which a prospective candidate for employment or a present employee is best suited.³ Particularly in times of labor shortage, when the best possible use must be made of all available man-power, it

² The top administrative jobs are likely to be filled on the basis of criteria other than those afforded by the tests. But the immediate direction of other workers, including the assignment of specific jobs to particular persons, instructing them concerning procedures, and overseeing their work, demands a particular combination of administrative capacity and social skill which is theoretically susceptible to measurement and for which tests of moderate usefulness are now available.

³ It should hardly be necessary to point out that candidates of foreign birth or those with little formal education should always be given a nonlanguage test rather than one that makes much demand upon understanding of English or reading ability, provided, of course, that the job for which they are applying does not require skill along these lines.

is likely to be wasteful of time and energy to proceed at once to tests of aptitude for highly specialized jobs. Better to begin with a determination of the subject's general level of ability and to go from there to a measure of his aptitude for one or another of the broad fields indicated under (2) above. With this basic knowledge as a starting point, more specific tests for types of work within the field for which he appears to show his greatest ability can follow.

Suppose, for example, that a firm has advertised for twenty-five machinists. Fifty men apply. Of these, six are rejected at once because of physical defects obviously unfitting them for the work or because of insufficient training or experience. An intelligence test eliminates four more whose scores are definitely below the critical score⁴ found necessary for success on the kind of work involved.

The remaining forty are then given a test of mechanical ability, or, if time and funds permit, a battery of such tests, inasmuch as no one of the many tests called by that name can be said to have the general meaning attributed to tests of general intelligence, nor are they equally free from the effects of previous experience. From the standpoint of industrial selection, however, the latter point is of less importance than it would be for the prediction of vocational aptitudes, although it may result at times in the choice of experienced workers of mediocre ability whose capacity for improvement beyond their present level is small in preference to those of greater potential capacity but little actual experience.

In the case we are now considering, comprehension of mechanical principles rather than mere dexterity of hand is required. A good combination is a battery composed of the Minnesota Paper Form Board Test (Likert and Quasha, 1934), which is not dependent upon mechanical experience, together with one or more of such tests as the Detroit

⁴ The critical score method is much used, not only in industrial selection but in the allied fields of selecting candidates for admission to universities or trades schools. The procedure is as follows: A large group of individuals is given the test in question at the time of their application for the job under consideration or for admission to the school. Selection is made, however, without reference to the test scores but according to whatever principle has traditionally been used, such as the judgment of the employment manager or the high school record. After sufficient time has elapsed to permit the subjects to demonstrate what they can do (say six months or a year later), their records are examined and compared with the test scores made at the time of entrance. If the test used has any value for the purpose at hand, it will usually be found that even though the relationship between score and performance is only moderately high, it will still be possible to select a minimum score below which the chances of success are so poor that they may safely be disregarded. This point is known as the "critical score value." The critical score is not necessarily a fixed point. It may be set at a higher level when the selection ratio (the ratio between the number of applicants to be chosen and the number available) is low and the standards for selection can therefore be more rigid. No advantage is to be gained, however, by moving it downward to a point so low that failure is well-nigh certain.

Mechanical Aptitudes Examination (Baker, 1929), the Bennett Test of Mechanical Comprehension (1940), or the O'Rourke Mechanical Aptitude Test (1937), all of which deal with the uses and operation of pictured parts of mechanical apparatus, such as pulleys and belts, cog wheels and levers, nuts and bolts. All of these paper-and-pencil tests may be given to the entire group at one time. The Paper Form Board Test has perhaps been more widely used in vocational guidance and in the selection of workers for positions requiring quick and accurate perception of spatial relations than any other single measure of its kind. Each item of this test consists of a picture of a plane figure which can be constructed by combining certain ones of a number of pictured smaller parts. The subject is required to indicate which parts he would choose. Selected items from some of these tests are illustrated in Figures 37, 38, and 39.

With the accumulation of data, multiple-correlation formulas may be worked out to indicate the best weights to be given to these tests and to others which may be substituted for or added to the battery for purposes of prediction. Until such information has been secured, the tests may be weighted in accordance with psychological judgment based upon knowledge of the requirements of the particular situation. The intelligence test may be included in the battery or may be used only as a screening device. Generally speaking, intelligence above a certain critical and necessary level is of relatively little importance for success in mechanical operations of a simple repetitive nature, but becomes an increasingly significant factor in success as the complexity of the tasks to be performed and the amount of independent judgment required for performing them are increased.

Returning to the instance in question, repeated experiments have indicated that by choosing from the forty candidates the required twenty-five who make the highest scores on the two or more tests designed to measure mechanical ability, a considerable improvement may be gained over methods which depend upon the judgment of the employment manager together with such information as is likely to be obtainable concerning previous employment records and the like. This does not, of course, mean that the tests will invariably predict success more effectively than the more traditional methods. Mistakes will be made in either case. But the *proportion* of successful choices made on the basis of the tests is likely to be greater than that based upon brief interviews and inspection of records. And when the number of men on the pay roll is large, the saving of expense brought about by the lowered turnover becomes a considerable item, even though some errors in selection will inevitably occur.

It may happen that when the scores on the mechanical ability tests

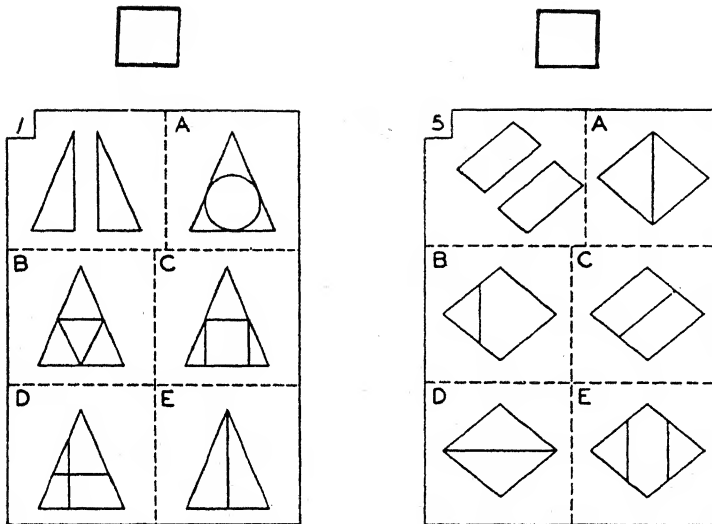


FIG. 37. TWO ITEMS FROM THE MINNESOTA PAPER FORM BOARD TEST. The subject is required to find the figure that can be made by combining the two parts shown in the upper left corner and to write its letter in the small square above. (Reproduced by permission of the senior author, Dr. Rensis Likert, and the Psychological Corporation of New York, publishers.)

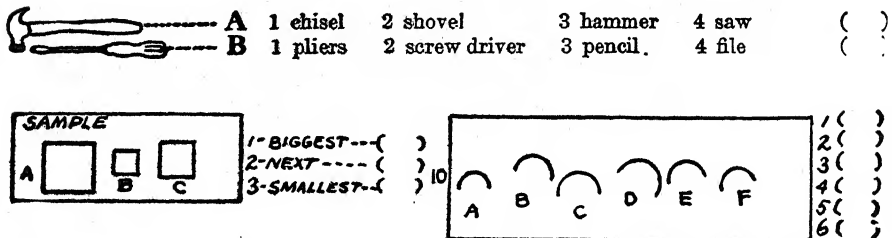
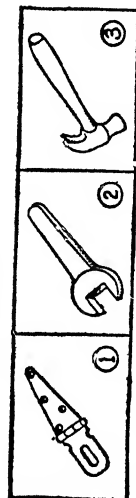


FIG. 38. TWO ITEMS FROM THE DETROIT MECHANICAL APTITUDES EXAMINATION. The Detroit Mechanical Aptitudes Examination includes eight sub-tests as follows: a picture vocabulary test, a test of size-perception, a test of precision of hand-movements, an arithmetic test, a test of picture arrangement, a test of mechanical information, a test of mechanical reasoning and a matching test. Sample items from the two first named are shown above. (Reproduced by permission of the author, Dr. Harry J. Baker, and of the Public School Publishing Company, Bloomington, Illinois.)

Each of the three pictures marked with a number is **used with** a picture at the right marked with a letter. Look at the picture marked 1. Then look at the pictures marked A, B, and C and decide which is **used with** 1. Write the letter of the picture which goes with 1, on the line marked 1 at the right of the pictures. Then find the picture that is **used with** picture 2, and write the letter of that picture after 2 on the line at the right. The first sample is done correctly. Picture C is **used with** picture 1, so "C" is written after 1 on the line at the right. B is **used with** 2, so write "B" on the line at the right AFTER 2. "Nail," marked A, is used with "hammer" marked 3, so write "A" AFTER 3 ON THE LINE AT THE RIGHT.

NUMBERED



LETTERED

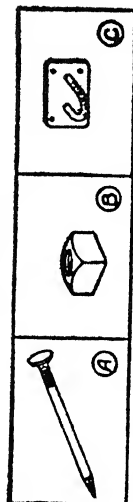


Fig. 1.

(Write answers here)

C

1. _____

2. _____

3. _____

Under each set of pictures you will find some questions. In **each** square at the right of the questions, write a number or a letter to show which tools you would use to do what is asked. Where there are two squares, be sure to write a number or a letter in **each** square. Pictures 3 and A show what is used to fasten a board to a box, so 3 and "A" are written in the squares at the right after question 1. Picture 2 is the correct answer for question 2. Pictures 1 and C are the correct answers for question 3, so write 1 in the first square and "C" in the second square after question 3.

In each square at the right of the questions below, write a number or a letter to show which tool in Figure 1 you would use:

1. To fasten a board to a box. . .
2. To tighten a nut. . .
3. To fasten a door so as to use a padlock.

**(Write an answer
in EACH square)**

1	2	
3		

FIG. 39. TWO ITEMS FROM THE O'ROURKE MECHANICAL APTITUDE TEST.

are compared with those on the intelligence test, some cases will be found for whom the latter scores are exceptionally high, even though these men are rated among the fifteen who fail to qualify for the machinist positions. They may well be considered for some other line of work. Perhaps additional salesmen are needed. No very good single test for the selection of salesmen is available, but the key for "persuasive" interests of the Kuder Preference Record and the three keys for measuring interests in special fields of salesmanship—life insurance, real estate, and sales management—of the Strong Vocational Interest Test described in a previous chapter should provide some cues.

For the selection of workers to fill the many jobs in which speed and dexterity of hand and finger movements rather than understanding of mechanical principles are required, such tests as the O'Connor Finger Dexterity Test, which requires the subject to fill each of 100 holes in a metal plate with three small brass pins, is often used. This apparatus is not unlike the pegboards, beloved of kindergarten children, except that the pins are smaller (.072 inches in diameter) and that three are placed in each hole. The score is the time required to fill the board. The self-correlation of this test is very high, well over $+.95$ for unselected subjects of similar age and sex, but its highly specific character means that it can measure only a very limited aspect of the skill required for most types of factory work. It is better used as a part of a battery of tests rather than singly. The apparatus can be obtained from the C. H. Stoelting Company of Chicago. Other tests used for the measurement of rather simple manipulative skills are the Minnesota Rate of Manipulation Test (Green, *et al.*, 1933), which measures reaction time for hand and finger movements; the Purdue Hand Precision Test in which the subject is required to punch a stylus into holes one-half inch in diameter that are uncovered by a rotating shutter at the rate of one hole per half second, without touching the side of a hole or getting the stylus caught in the shutter. The Hayes pegboard (1932) is in some ways similar to the O'Connor Finger Dexterity Test previously described but involves hand and arm coordination as well. Many others are described in such books as Bingham's *Aptitudes and aptitude testing* (1937), *A manual of selected occupational ability tests* by Green, Berman, Paterson, and Trabue (1933), *Industrial psychology* by Tiffin (1942, Revised 1947), and the *Encyclopedia of vocational guidance*, edited by Kaplan (1948).

The distinction between aptitude tests and tests designed to measure the degree of knowledge or skill already acquired is an important one for the industrial psychologist to bear in mind. When the work to be done is of a relatively simple kind which the worker can master in a few hours, aptitude for the job is the main thing to be considered, and the

more completely the test used is freed from the influences of previous experience, the better, other things being equal, is it likely to serve the purposes of selection. When, however, the job is one for which a considerable amount of training is needed before the employee becomes of much practical value to the employer, aptitude tests are of less immediate value than tests of achievement or skill already acquired. The majority of clerical jobs are of the latter kind. Stenographers and typists, bookkeepers, and even filing clerks must undergo a considerable period of training in order to learn their jobs. For this reason, tests of skill in the particular kind of work for which they are to be employed, of which a number are described in the references previously cited, will be used in preference to the aptitude tests used in vocational guidance. However, particularly in the case of young employees on low-level clerical jobs, vocational promise as indicated by the aptitude tests may also be taken into account.

The personality characteristics of the worker count for as much as or more than his actual skill. No matter how able a workman may be, if he is quarrelsome, uncooperative, undependable in his habits, and a continual source of friction among his fellow workmen, he is likely to be more of a handicap than an asset. In contrast, the man of moderate ability who is steady and dependable, whose relations toward his companions and supervisors are easy and friendly, who has a good sense of humor and is not readily upset by minor irritations will do more to keep the plant machinery running smoothly and efficiently than many another who outstrips him in work output but who is unfriendly and irritating in his personal relationships.

The use of the Minnesota Multiphasic Personality Inventory as a screening device for prospective employees should at least help to identify in advance most persons whose maladjustment has reached a stage which is likely to make them serious troublemakers among other employees. It is quite possible that marked deviations in the opposite or "desirable" direction on the various keys of this test might provide valuable clues as to the types of work which persons showing the particular temperamental patterns thereby indicated are likely to find most satisfying. Up to the present time, however, little has been done to determine the significance of "good" scores on the various keys of this test. Investigation of this question might prove highly worth while.

Other personality inventories have been used as partial bases for industrial selection and the placement of workers, but on the whole with only a very moderate degree of success. Tiffin (1942, page 116) reports the successful use of the Humm-Wadsworth Temperament Scale (1940) in the selection of employees at the Lockheed Aircraft Factories at Bur-

bank, California, and Martin (1944) describes the successful use of the "cooperativeness" component of the Guilford-Martin Personnel Inventory in locating the prospective troublemakers in an industrial plant.

As Tiffin has pointed out, however, one of the major difficulties with all inventories of this kind is that applicants for a job are under strong incentive to make a good impression and are likely to reply to the questions in terms of the answers they judge to be desirable, rather than in terms of truthful self-report. This hypothesis has been tested by having college students fill out such questionnaires first with the instruction to reply to each question as truthfully as possible, and a second time with the instruction to imagine that they were applying for a job and to answer the questions as they would under those circumstances. The shift of scores toward the desirable end of the scale was very marked. Whether this tendency renders such scales entirely invalid for industrial placement or whether all that is necessitated is a change in the position of the critical score is not known with certainty, but unless the amount of the shift is approximately uniform for all subjects—which seems improbable—it unquestionably sets definite limits to the amount of confidence that can be placed in the findings.

THE VALIDATION OF TESTS FOR INDUSTRIAL PURPOSES

Because experience affects the scores on most tests used for the selection of workers, scores made by such persons cannot be used as normative standards with which to compare those made by candidates at the time they apply for work. For this reason and others it is better to begin by testing all candidates at the time of application but without using the test scores as a basis for selection, which will be made in accordance with the traditional methods. After a sufficient period has elapsed for the determination of success on the job, the following comparisons should be made:

1. Mean score of rejected vs. accepted candidates. In connection with other methods to be used, this will afford some evidence as to the validity of the methods of employee selection previously used.
2. Mean score of candidates who were discharged or who left work because of some real or fancied grievance during the trial period, compared with that of those who remained on the job.
3. Mean scores of candidates whose work, at the time of the checkup, was rated by supervisors as "good," "average," or "poor."
4. Comparison of scores with daily or hourly output when the latter measures are available.

5. Comparison of scores with quality of output as measured by amount of work rejected by inspectors or other available evidence.

6. If possible, some measure of worker satisfaction should be obtained, on the assumption that those who are suitably placed will, on the whole, be more contented and have fewer complaints to make than the occupational misfits.

7. If personality tests of any kind were included in the original battery, some evidence concerning the relationship of the workers to their shopmates and supervisors, and their apparent role as troublemakers or as trouble preventers should be secured. Inasmuch as only rather outstanding cases are likely to be noted, foremen may simply be asked to name the two or three men most noticeable for each of the two qualities mentioned.

From the data obtained in this way, tests which seem to have been most effective for the employment of workers and their placement on particular jobs may be selected and combined into batteries in which weights have been determined by multiple correlation. The usefulness of these batteries should then be ascertained in terms of the improvement over previous selective methods that is obtained when future workers are selected on the basis of their test scores rather than by the methods previously used.

That industrial testing is a profitable investment for a firm employing large numbers of workmen has been amply demonstrated, assuming, of course, that the testing program is in the hands of a competent psychologist. Miracles must not be expected, for no test is infallible, and psychologists, like other human beings, will inevitably make errors. But for the industrialist the question resolves itself into a simple one of dollars and cents. Is the saving in overhead brought about in this way greater or less than the cost of the testing program? Most firms who have investigated the matter have found the testing program worth continuing.

THE QUALIFICATIONS OF THE INDUSTRIAL PSYCHOLOGIST

Success in the field of industrial psychology depends upon the degree of technical knowledge and skill possessed by the psychologist and upon his personal qualifications for a type of work in which social relations play so important a part. He must be able to command the respect of both shopmen and managers for his ability in his own particular field, and at the same time be able to meet them on their own level in everyday intercourse. He should be a fluent speaker, able to address groups on informal as well as formal occasions and to present his views in a clear and convincing manner. Since industrial psychology is a comparatively new field, he should recognize the need for demonstrating what it can do.

To this end he should know how to make full use of charts and lantern slides, and how to boil down his tables to a small number of salient figures, the significance of which can be quickly and easily grasped. Publication of his results in trade papers as well as in professional journals will help to bring the possibilities of a sound program of psychological testing before the eyes of those who are in a position to make use of such a service. In all this he must, of course, be careful not to overstate his case. Better for him to err on the side of conservatism than to lay himself open to the charge of charlatanism. He must show not only what tests are capable of doing for an industrial concern but also what they have failed to do—the percentage of failures to be expected as well as the proportion of successful placements. He must be careful to give full credit to the more traditional methods of employee selection, and to avoid giving the impression that just because the new methods have been found to work better than the old, the latter should be regarded as worthless.

His own technical equipment should include not only a thorough grounding in general psychological principles but some experience in laboratory psychology of the more conventional sort, as well as intensive training in the field of tests and measurements. His statistical training should extend far beyond a knowledge of formulas and ways of using them. He must understand the principles upon which these formulas are based and be able to apply them to the concrete situations with which he deals. Far too many industrial psychologists feel that they have learned all that it is necessary to know about a test when they have ascertained and reported its “reliability” for a group of unstated composition, together with its correlation with a “criterion” of unknown worth for the purpose at hand. It sometimes appears as if much of the training of the prospective industrial psychologist is of a kind that emphasizes processes rather than principles, and exalts figures without much regard to their basic significance.

This, needless to say, is by no means universally true. There are many well-trained persons whose work in industrial psychology is contributing much, not only to the firms by whom they are employed but to the broader area of social psychology and human adjustment. The average man spends almost a fourth of his adult life on his job. Whether he finds it a source of personal satisfaction or an irritation and a strain, whether the material contribution which he thereby makes to society represents a large or a small part of what he is potentially capable of producing are not trivial questions. The national economy depends upon the work of the average citizen and it is with these men and women that the industrial psychologist has to deal.

Testing and Social Welfare

PSYCHOLOGY AND MEDICINE

Both the pediatricist and the family physician are frequently called upon for help in cases of mental subnormality. Distressed parents who find that their child is not developing normally turn to their doctor in the hope that some remediable physical cause may be responsible. Too often, in the past, if the child appeared to be in good physical condition the parents were assured that the backwardness would be outgrown in time, or "after he reaches puberty." That such improvement rarely occurs in fact was either unknown to many physicians of the old school or purposely ignored by them. Those with more recent training are better informed on the subject. Recognizing their own lack of specialized training in this area, they frequently call upon psychologists for aid in diagnosis and in prognosis. In some medical clinics a psychologist is included as a regular member of the staff. Elsewhere a kind of informal cooperative relation has been established between one or more private physicians and psychologists, with each referring cases to the other when need for advice is felt.

Practically all up-to-date mental hospitals now have a special department devoted to the psychological examination and treatment of patients. The work done involves much more than just a routine administration of intelligence tests, which was once regarded as the essential function of the clinical psychologist. The modern psychologist who deals with mentally abnormal patients approaches his task from a very different angle. He gives intelligence tests, to be sure, but less to attempt to ascertain the subject's present level of performance than to attempt to estimate, on the basis of the particular pattern of successes and failures together with data obtained from other sources, what his mental capacity was most likely to have been before the onset of the disease. From a comparison of the two estimates¹ some indication is provided of the extent of mental deterioration that has resulted from the disease.

¹ An intelligence test given to a mentally disordered person is not wholly comparable to one given to a person in a normal state of mental health. In the latter

The appraisal of the patients' mental capacities is, however, but one feature of the work of the psychologist attached to the staff of a mental hospital, and intelligence tests represent but a small part of his psychological equipment. His aim is the restoration of as many of the patients as possible to a state in which they can again function as normal members of society, and to that end he seeks constantly for such diagnostic instruments as will provide clues for therapeutic treatment. Much of his work will of necessity be experimental, for there is no other field in which research is more urgently needed, and in which the psychologist who is too timid or too lacking in scientific imagination to venture upon new paths will be of so little use. The methods and practical applications of abnormal psychology are by no means so clearly marked out as are those used by the educational psychologist attached to a public school system. There the conscientious follower of formal rules and principles but with relatively few ideas of his own can do a fairly satisfactory though uninspired piece of work. He will add little or nothing to the scientific knowledge already available, but if he has been thoroughly trained in the techniques of measurement worked out by his more gifted predecessors and is careful and accurate in his work, his appraisal of the aptitudes and skills of the rank and file of school children will for the most part be reasonably accurate. But no such clear-cut rules of procedure are available for the psychologist in the mental hos-

case, cooperation can usually be taken pretty much for granted. The attention can be secured without great difficulty and held for the time necessary to administer the test, nor is one likely to run afoul of unexpected emotionally toned associations. Thus the test may, in most cases, be looked upon as a valid sample of the subject's ability to perform tasks of the kind indicated. But in the case of the mentally diseased patient, such an assumption is far more hazardous. His attention is likely to waver from time to time, his attitude toward the test may be one of suspicion, indifference, or resentment. If his cooperation cannot be secured at all it is, of course, futile to attempt a test. But among those who appear to cooperate, interest and effort may fluctuate greatly from moment to moment. Although some would argue that this very fact is a significant part of the intellectual picture, that the patient's inability to bring all his powers to bear upon the solution of a problem is an essential feature of his mental state, they overlook the very important point that this variability of attention and effort decreases the dependability of the test score. If its only effect were to lower it, that would be a different matter, and the argument just mentioned would be sound. This, however, is not the case. The increase in the error of measurement is likely to be of greater consequence than the average loss from lack of ability to focus the attention, since it means a corresponding reduction in the amount of confidence that can be placed in the results. For this reason several tests should be given on different occasions in order to make up in part for the increase in the sampling error. Even when this is done, however, the experienced clinician is likely to use the results as a basis for *estimating* the subject's mental level, instead of regarding them as a valid sample to be interpreted by comparison with a table of normative standards. In making such an estimate, he will be guided both by the quantitative results of the test, including the pattern of successes and failures, and by the patient's manner of attacking the tasks and his behavior while working on them.

pital, in spite of the tremendous amount of research work that has been done in that area. He gives tests according to standardized methods but he must always be on the alert for behavioral signs which may alter their meaning for particular patients. He must compare these observations with other data in the hope that as their meaning becomes understood they may be added to the series of diagnostic signs by means of which an analysis of the patient's difficulties and a prognosis as to his chances of recovery are attempted.

The clinical psychologist in the mental hospital should therefore be first of all a qualified research worker. He should be thoroughly conversant with the tests and other diagnostic tools for which some degree of scientific validity and clinical usefulness has already been demonstrated. He should know their possibilities and recognize their limitations and thus be in a position to use them for what they may be worth. Through the current experimental literature and attendance at professional meetings as well as by such direct personal acquaintances as he may form, he should keep in close touch with the work of his colleagues in other institutions. He should maintain an open mind with respect to the claims made for new theories and methods, along with an attitude of healthy skepticism which refuses to permit the acceptance of these claims until they have been substantiated by objective evidence.

One of the ways by which his routine work may be made to contribute both to the improvement of his own diagnostic and therapeutic procedures and to the advancement of scientific knowledge in the field is the maintenance of a cross-indexed file which permits the comparison of a particular method or diagnostic sign with the characteristics of each of the patients for whom data on this head have been obtained. Conversely, the system should also permit the comparison of a particular patient with all the others either in terms of a particular pattern or type of performance or in terms of some single attribute or measure. The use of Hollerith cards will greatly facilitate research carried on by this method.

Suppose, for example, one wished to check the statement made by Bochner and Halpern (1942) that schizophrenic patients give very few movement responses on the Rorschach test. By means of the cross index, two tests of this hypothesis can be made. The first involves a comparison of the average number of movement responses given by patients diagnosed as schizophrenic with the corresponding figure for other clinical groups. By means of the *t* test, the level of confidence that may be placed in the findings can be determined. But since the diagnosis of schizophrenia by such means as are now available is by no means a wholly objective matter, a second check on the hypothesis may be made by working on the other regression line. By selecting all the cases for

whom the number of *M* responses fell in the lowest quartile or, if the number of cases is sufficiently large, the lowest decile of the total distribution, not only may the formal diagnosis of the clinical syndrome (schizophrenic or nonschizophrenic) be taken into account, but a special study of the cases that do not conform to the hypothesis can be made. It may be concluded in some cases that the original classification was erroneous, in which case the Rorschach sign would take on added significance. Or it may be found that the number of cases that conform to the rule is no greater than the number that run contrary to it, or that the difference is no greater than might reasonably be expected by chance. In such cases, further use of the sign for diagnostic purposes would not be profitable.

Mental testing in hospitals for patients suffering from various chronic ailments or from conditions requiring a long period of hospitalization has also proved its worth. In such cases the measurement of morale, of attitudes toward the self and the outside world, of anxieties, worries, resentment, and other emotional states may be useful. Here the various personality inventories come into their own, not so much from the standpoint of numerical scores but as a basis for personal interviews and counseling. Tests of interests and of vocational aptitude provide a partial basis for guiding occupational therapy for such patients as are able to profit by it, as well as for vocational guidance after the patient has left the hospital, especially in the case of those patients forced into a new and different type of occupation by partial disablement resulting from the disease.

THE USE OF MENTAL TESTS BY SOCIAL WELFARE AGENCIES

For many years social welfare agencies have wrestled with the problem of the mentally defective client. Over and over again they have doled out financial help, found jobs for him which have been as promptly lost, given fruitless advice, paid calls, and held staff conferences.² These agencies are now taking a more realistic view of such

² The following is an example personally known to me. The moron father of a family consisting of a wife and four living children was taken ill and sent to the free ward of the city hospital. A social worker called at the home and was gratefully received by the wife, who told of their destitute condition. Her soft voice and gentle manner made a most favorable impression on the worker, who felt that if the family could be helped over the immediate emergency until the father was once more able to work, and if a job could then be found for him, all would go well. She made an inventory of their needs, had coal and food sent in, and clothing for the children.

situations and are calling upon mental examiners to help them to differentiate between those clients who can be helped to help themselves, and those for whom the likelihood of successful management of their own affairs without aid from others is so remote that continued supervision is likely to be the only feasible solution for them.

Not only in the diagnosis of mental deficiency but in many other questions with which the social worker has to deal can the psychological examiner be helpful. Modern social service is directed toward rehabilitation of the client, and this frequently means a combination of the roles of vocational adviser, personal counselor, and father confessor in addition to the more traditional service of giving material aid to the needy. There is accordingly a growing realization on the part of directors of schools of social service that prospective social workers should learn enough about the general principles and methods of mental measurement to enable them to make an intelligent use of the results of such measurements as reported to them by qualified examiners. In addition, some schools are encouraging their students to learn to administer a few of the more simple tests of this kind.

The various organizations for children and youth, such as the Scouts, the Campfire Girls, the Hi-Y Clubs, and the junior Y.M.C.A. and Y.W.C.A. groups as well as summer camps for children, not infrequently find their work interfered with by the presence in the group of some youngster who acts as a disturbing influence. The more progressive leaders are beginning to realize that the clinical psychologist can often throw light on causes of such behavior and so help in resolving the difficulty.

Since two of these were without shoes to wear to school, she entrusted the mother with money to buy the shoes.

A few days later the woman called at the welfare office and asked to see the social worker. With a beaming face she offered a parcel which, when opened, was found to contain a large cabinet photograph of the entire family. The amazed and discomfited worker asked where the money for the picture had been obtained. "Oh, you know. You give it to me," was the ready answer. "But," protested the worker, "that money wasn't to be spent for a picture. That was to buy shoes for the children, so they could go to school." "Yes," admitted the woman, "I know you said so. But I got to thinking. Sometimes kids, they die. And by and by you can't remember 'em good no more. So I thought, well, they got these good clothes you brought 'em an' they all look so nice I better get their pictures taken while I got 'em."

This is a good example of moron reasoning. The woman loved her children and wanted to keep them with her, but she had learned by bitter experience that "sometimes kids, they die." Six of hers had already gone that way. That time had blurred their images in her mind was an added grief; the picture would be a reminder if the remaining four should also be lost. The futility of attempting to help such people to manage their own affairs with ordinary prudence is apparent.

THE PROBLEM OF THE UNMARRIED MOTHER

Organizations dealing with the unmarried mother have a threefold problem to consider. First of all there is the disposition of the child. Shall it be left in the care of the mother and her family, given temporary placement in a foster home, or made immediately available for adoption? Regardless of which plan is adopted, it becomes necessary to make a careful study of the child and his antecedents with a view to making the best possible estimate of his mental and physical potentialities. It is equally necessary to investigate the home in which he is to be placed, not only as to its physical aspects but even more carefully with respect to the personal characteristics of the members of the family. The selection of a home should be made with reference to the characteristics of the child, as far as these can be judged from the available data. Although intelligence tests for young infants are of little value for predicting their later development, something can be learned by studying their ancestry, and in this connection tests given to the mothers provide important evidence. In evaluating the homes, the use of the Leahy scale previously mentioned or of the Williams scale for measuring home environment (1916) and of one or more of the better scales for measuring the foster parents' attitudes will provide some objective evidence to supplement that obtained through unstandardized observation and interviews. Since many tragic disappointments have occurred from the placing of unsuitable children in adoptive homes or from the selection of unsuitable homes for the placing of children, the agencies concerned with child placement should be alert to take advantage of every diagnostic aid that may help them to wise decisions.

The second point to be considered is that of the girl herself. As a rule these mothers are still young, with the greater part of their lives ahead of them. They must be helped to make the most of these lives, to regain their self-respect, and make for themselves a secure place in the social and industrial world. A thoroughgoing appraisal of their intellectual and personal assets will help to show their vocational potentialities and in many cases will reveal points of special weakness in the make-up of their personalities which should be regarded as danger signals, warning against certain environmental hazards which are likely to be greater than these girls can easily withstand.

Finally, there is the question of the girl's family and of her relation to it. If her people consider that she has permanently disgraced herself and them, if they treat her as a criminal or an outcast or attempt to keep her under such close surveillance that she rebels and perhaps leaves home with no well-conceived plan as to where she is to go or what she

is to do,³ not only the girl's future but the happiness and well-being of the other members of the family is jeopardized. One of the important tasks of the welfare agency therefore becomes that of helping the family to a better understanding of the girl and her difficulties, and of reinstating her in its affection. If this cannot be done, an attempt should be made to find a family substitute in the form of some older person or persons to whom she can turn for advice and help and who will provide an emotional anchor to steady her against adverse winds. While the direct use of tests is not likely to be practicable here, a tactful interpretation of what they have revealed about the girl herself may sometimes help to make her more understandable to those who deal with her.

THE PREVENTION AND TREATMENT OF JUVENILE DELINQUENCY

In the literature on juvenile delinquency, three somewhat divergent points of view can be noted. First there is the environmentalist position exemplified by Professor Clifford R. Shaw of the University of Chicago. By mapping the location of the homes of boys brought before the Chicago Juvenile Court over a period of years, Shaw (1929) was able to show that there are certain districts of that city that may properly be designated as "delinquency areas," since the proportionate number of juvenile delinquents whose homes are in those regions vastly exceeds that to be expected on the basis of a random distribution of cases according to the city's population. Similar conditions have been found in a number of other cities studied by Shaw's method. These areas are notable for the existence of many more than the usual number of conditions likely to have an adverse effect upon the behavior of children and youths, such as saloons, gambling houses, and "hangouts" for tramps and vagrants as well as for the almost total lack of playgrounds or other provision for healthful recreation.

That the typical young delinquent makes a rather poor showing on standard intelligence tests has repeatedly been established. Although it is now known that the exceedingly high proportions classed as "feeble-minded" by the early investigators of the problem were in large part due to errors in the standardization of the tests used, more recent studies using modern tests have not erased the intelligence differential. Glueck and Glueck (1934) compared the Stanford-Binet IQ's of 979 young delinquents with those of over 3000 Massachusetts school children of

³ Many unmarried mothers are so apprehensive of the way their misdemeanor is likely to be received at home that they anticipate the difficulty by running away before the baby is born.

similar age. More than 30 per cent of the delinquents, but only slightly over 7 per cent of the school children had IQ's below 80. These figures do not differ greatly from those in the more recent report by Merrill (1947), in which cases passing through a juvenile court in California are compared with those used for the original standardization of the 1937 Stanford-Binet, which was also used for the delinquents in this study. Slightly over 25 per cent of the delinquents and 8.2 per cent of the standardization group earned IQ's below 80. Other modern reports agree fairly closely with these figures. In round numbers it may be said that at least 10 per cent of the apprehended juvenile delinquents⁴ of today are of such low mentality that attempts to make them fit for useful life in society are almost certain to be wasted. In an institution under proper supervision, however, many of them can be partially self-supporting.

Even if 10 per cent of apprehended delinquents are mentally defective, there remain the 90 per cent who are intellectually capable—or should be—of restoration to normal life in society. For the 10 per cent, the main problem is that of identification, though it is true that psychologists can often help in the sometimes difficult task of convincing the court and the child's family that permanent commitment to an institution is the only wise solution for these cases. The real problem, however, has to do with those whom it should be possible to restore to society.

The great bulk of these cases will usually be between the ages of twelve and eighteen years. Of the 300 cases studied by Merrill, approximately 90 per cent fell within that range, and almost 70 per cent were from fourteen to seventeen years of age. Juvenile delinquency is therefore almost entirely an adolescent manifestation, though its roots extend backward into childhood and its consequences reach forward into adult life. The Gluecks found that 88 per cent of their cases were recidivists within five years after the original study,⁵ but after fifteen years the proportion who had been convicted of further offense during the last five years of that period had dropped to 66 per cent. Some of the earlier offenses, so the Gluecks believe, could be attributed to boyish love of adventure and youthful exuberance, rather than to more deep-lying conflicts or basic personality defects. With the maturation of the personality, and the development of planfulness, foresight, and self-control that are its normal accompaniments, criminal behavior no longer

⁴ Particularly in questions of intelligence, it is always necessary to distinguish between apprehended delinquents and delinquents in general, since those of higher mentality are in all probability more successful in eluding the police.

⁵ In the sense that they had been guilty of at least one further offense during that period.

occurred. The recidivists, in his opinion, were in most cases individuals in whom this normal maturation of personality did not take place. Although adult in years, their personal-social reactions continued to resemble those of adolescents.

As she correctly reports, the considerably smaller proportion of recidivists reported by Merrill (48 per cent during the five years immediately following the first referral), than that found by the Gluecks, arises from the fact that the Glueck's study dealt only with the more serious or difficult cases which the judge felt incapable of handling by himself and therefore referred to the Judge Baker Foundation for further study. Merrill's cases, on the other hand, include all those passing through the juvenile court in question during a specified period. Nevertheless, even if the proportion of recidivists is no higher than 48 per cent, the need for doing something about them is obvious. Merrill also feels that recidivism alone is not a fair indication that treatment has been unsuccessful. Her estimate that at least 80 per cent of the delinquents whom she studied were at least fairly well adjusted by the end of the five-year period has a more hopeful sound, but is based on rather subjective evidence. One could wish, too, that her reports of treatment had been more adequate. For the most part she describes only the diagnostic methods used, the distribution of scores made by the delinquents and the nondelinquents with whom they were compared, together with critical ratios. There are also a number of brief case reports and some statistics on home conditions, which, in conformity with the findings of most other investigators of delinquent behavior, are less favorable for the delinquents than for the average child.

It is highly unfortunate that up to the present time psychological studies of the young delinquent have been so preoccupied with determining what the typical delinquent is like that they have given almost no attention to the possibility of carrying out truly experimental investigations on methods of treatment or of prevention. Nevertheless, large sums of money have been spent for both purposes. Volumes have been written describing these projects. Hundreds of published case reports have described in colorful and moving terms the sordid and frustrating environment in which many of these cases have lived, the difficulties they have encountered, and the lack of understanding with which they have been treated. All this has undoubtedly tended to awaken public sympathy for the offender. "There, but for the grace of God, go I" is the feeling engendered by these reports. The inculcation of such an attitude may be desirable, but it takes us no closer to the heart of the problem. Merrill, in common with many other psychologists, has ad-

ministered a long list of standard tests to her group of subjects and has collected much information concerning their home background and school history, their interests, amusements, and companions. Like others who have investigated the same problem, she has found a large incidence of broken homes, generally unfavorable home conditions, unwise parental management, poor school progress, and many other adverse facts, all of which have been carefully summarized with the results presented in tabular or graphic form. She has also made some attempt to study the relationship of various factors to delinquency by comparing their frequency in delinquents and in matched controls, in recidivists and in nonrecidivists, and so on. But this is all *ex post facto* reasoning. What is needed is an experimental approach to the problem. Granted that under existing social conditions and legal restrictions it is not always easy to set up the experimental conditions that may be desired, some problems may nevertheless be attacked in this way. Such techniques as play therapy, nondirective counseling, psychoanalysis, placement in foster homes, probation in charge of a sympathetic and interested "Big Brother," and many others have their ardent supporters. Selected instances in which apparent success has attended the use of each method are cited as evidence of its superior advantages. But these are not experiments.

The psychologist of today has many resources at his command that his predecessors lacked. He has a better concept of experimental design. Imperfect as many of his tests are, he nevertheless has a sufficient repertoire for which the meaning and dependability have been well enough established to make these tests serviceable tools of research when used with understanding and discretion. The problems of juvenile delinquency and of adult crime are matters that affect all of us, but the methods used for ascertaining how they may be ameliorated have for the most part been formalized and unimaginative, centering upon a tabulation of the frequency of certain background conditions existing for a specified group of delinquents, and of certain traits of the delinquents themselves as indicated by their test performances. Such treatment as is given is intuitive rather than experimental. The excuse for this is that such great individual variations exist from one case to another that scientific inquiry of the usual sort becomes impractical, that clinical experience and psychological or psychiatric insight is the only possible guide. With inadequate records and only very loose criteria for determining the success of the methods used, it is not surprising that so little inroad has been made upon the problem of reducing crime and delinquency.

The time has come for another line of attack. Suppose that among

the first offenders⁶ brought to a juvenile court, two or more groups of cases are chosen according to some predetermined plan of selection. Case histories are obtained for each subject and a fairly comprehensive series of psychological tests and measurements and a physical examination are given. Each group is thereafter subjected to a specific type of therapeutic treatment. For example, nondirective counseling might be used with one, psychoanalysis with another, probation in charge of one or more carefully selected Big Brothers with a third. At the end of a specified period of treatment which should be the same for all members of each group, a checkup by means of a series of attitude tests, personality inventories, or whatever means seems suitable should be made, and at the end of a longer period, say five years,⁷ the extent of recidivism in each could be compared. By a further comparison of the distribution of successes and failures in the various groups with the data secured when the investigation was undertaken, hypotheses could be formulated with respect to the type of treatment best suited to various types of cases, and these hypotheses could then be tested by later experiment. Similar experiments could be set up in reformatories and in penitentiaries, with improvement in behavior inside the institution as a preliminary check and absence of recidivism after discharge as a final criterion. So long as our understanding of the most effective ways of rehabilitating those who have transgressed the laws of society rests upon unproven theories and fallible judgment, progress will of necessity be slow. The application of established rules of scientific investigation might do much to bring about a more intelligently oriented plan of attack upon this vital area of social welfare.

⁶ "First offenders," as the term is here used, refers only to court appearances. Many if not most of the subjects will have been guilty of previous minor delinquencies for which they were not apprehended.

⁷ The use of the five-year period would have the advantage of making the results directly comparable with those of other studies.

Testing the Armed Forces

THE USE OF TESTS IN WORLD WAR I

When the United States entered the war in April, 1917, there were few who seriously entertained the idea that mental measurement could find a useful place in military technique. It is questionable whether even the psychologists who comprised the committee that developed the justly famous Army Alpha and Army Beta tests had a real conception of the magnitude of the structure for which they were laying the foundations.

The psychological studies of human abilities made at that time demonstrated several things of vast importance for the future of testing. In the first place it was shown that the mental test is much more than a device for helping to identify the feeble-minded, and that normality is not an all-or-none phenomenon. That there are degrees of mental ability among the rank and file of mankind was generally recognized, but the extent of these differences was rarely appreciated. The idea that a forty-minute test would reveal more about the intelligence of a given person than could usually be had through weeks of acquaintance was so radical that few indeed were willing to accept it.

The second important fact brought out by the army testing program is that mental testing is not necessarily the costly individual procedure which had been the only known method before the beginning of the war. Methods of testing large numbers of persons at one time would soon have been worked out anyway, for the Army Alpha was modeled after a group test upon which Otis was working at that time. But it would in all probability have required a number of years for group testing to gain the prestige that it received as a result of less than two years' work in the army, during which time the tests were given to more than a million men and the results were analyzed in a far more elaborate and complete manner than any private individual would ever be likely to duplicate.

Finally, the value of the tests for the practical classification of men

was convincingly shown. The military authorities were quickly brought to see their worth, and no sooner were the tests released for civilian use after the close of the war than they were eagerly adopted for use in schools and colleges, clinics and private agencies, business firms and industrial plants. Other tests modeled after the army plan were soon devised for use both with children and with adults. Within the short space of two or three years, group testing was put on a firm basis.

Although many other types of tests were tried, only three of the group tests used in World War I can be said to have yielded even moderately successful results. Of these the Army Alpha, which is too well known to require description here, was by far the most dependable. The Army Beta, a nonverbal intelligence test used with illiterates, proved less valid for its purpose than the Alpha, but was still sufficiently discriminating to be worth using. Men earning questionable scores on this test were usually given an individual series of nonverbal tests as a check. The third test proved to be the progenitor of the modern psychoneurotic inventory. It was devised by R. S. Woodworth and, under the title of "Personal Data Sheet," was used with fair success as a screening device for locating men unfitted for war service because of temperamental difficulties. Many of the questions in this inventory have been included without essential change in modern devices of the same kind.

Some attempts were made to devise tests of special ability to be used in assigning men to various types of military duty. Information tests designed to test the subject's knowledge about a particular type of work proved to have some value. Mechanical assembly tests and a few other devices for appraising special abilities were also tried but were found to be only moderately useful. The great achievement of the psychologists in World War I was the group intelligence test.¹

PSYCHOLOGICAL EXAMINING IN WORLD WAR II

The psychologists responsible for the testing program in World War II came to their task under very different circumstances. In 1917, Yerkes and his colleagues had little in the way of precedent to go by. Less than a decade had elapsed since the first successful use of individual tests for measuring the intelligence of children. Although the newly

¹ A complete description of these tests with copies of the blanks used and directions for giving and scoring the tests may be found in *Army mental tests* by Yoakum and Yerkes (1920). For a more complete account of the results obtained see the official report edited by Yerkes and published as Volume XV of the *Memoirs of the National Academy of Sciences* (1921).

published Stanford-Binet included tests for the adult level, these had been used but little. In the spring of 1917, when work on the construction of the army tests was begun, almost nothing was known about methods of measuring the intelligence of the normal adult.

In contrast, the psychologists of World War II were confronted by a surfeit rather than by a deficit of testing devices, and by a volume of experimental literature on the topic with which even the most rapid reader could scarcely hope to keep abreast. They were accordingly less concerned with devising wholly new methods than with sifting those already available in order to find the ones which were or could be made suitable for their needs.

These needs, moreover, were much more specific and varied than had been the case in World War I. The change to a highly mechanized type of warfare made new demands upon the men engaged in it. Military organization became more complex. Individual roles were more highly specialized. The work of the twenty-odd years between the two wars had clarified the problem of group intelligence testing sufficiently to make the preparation of the Army General Classification Test used in World War II² a relatively simple matter. The two major problems that the psychologists of World War II had to meet were, first, to devise a series of aptitude tests for particular kinds of military duty, and, second, to work out more adequate means of identifying men who were temperamentally unfit for combat duty although they might be entirely capable of carrying on other types of military service.

Neither of these tasks was a wholly new one. Aptitude testing had been engaging the attention of industrial psychologists and vocational guidance workers for many years; there remained only the task of developing new tests for certain aptitudes (such as that of the bombardier) not required for civilian life, and improving others already available to make them better adapted for the sterner requirements of military service. Tests for selecting airplane pilots, for example, had

² The AGCT was made up of three subtests: a vocabulary test, a block-counting test in which the task was to determine the number of blocks in each of a series of pictured piles, and an arithmetic test. It was the major instrument used for the initial classification of inductees and has been given to well over 12,000,000 men. The test and the process of standardization are described in an article prepared by the staff of the Personnel Research Section of the Surgeon General's Office (1947). Scores were transmuted into standard units by a method similar to that of the T-score described in Chapter 13, but different constants were used. The mean was set at 100; the standard deviation at 20. The resultant scores thus bear some resemblance to the familiar IQ, but the spread of the scores about the mean is greater, with the result that less significance can be attached to very high or very low figures than for IQ's computed by the usual method or for the IQ Equivalents used by Goodenough and Maurer, though the difference is not so very great. An AGCT standard score of 140 would correspond to an IQ Equivalent of 134, an AGCT score of 90 to an IQ Equivalent of 93, etc.

been developed before the war and were intended primarily as screening devices for eliminating unpromising candidates for instruction in piloting commercial planes. This is a much simpler task than that of making a positive selection of men possessing the special attributes required for flying a bombing plane or a fighter plane. New tests for these purposes were therefore required as well as tests for many other special types of duty for which the requirements of military service are so different from those of civil life as to make the tests developed for use in the latter setting of little use in the armed forces.

The second major task of the army psychologists was that of determining the efficiency of available personality tests and tests of emotional stability for use in selecting men likely to break down under the stress of active military duty, or, on the positive side, for use in identifying those whose personality and temperament made them particularly well adapted for particular types of service. Many special studies along both lines were carried out. "Combat fatigue" was a term especially coined to express a combination of physical and nervous exhaustion very common among men who had been subjected to long and arduous experiences at the front. It was found, however, that such objective measures as the length of active combat duty were but slightly correlated either with the length of time required for recovery or with the apparent severity of the symptoms. The lack of correlation appeared to indicate that the characteristics of the man, his ability to withstand nervous and emotional strain, were quite as important as were the external conditions to which he was subjected, a finding, of course, entirely in accordance with observations made in the preceding war as well as with everyday experience. Many attempts were made to predict susceptibility to combat fatigue and to devise psychological methods for treating it, but with only very moderate success.

In addition to the major tasks of classifying men in accordance with their general intellectual ability and their special aptitudes for particular types of service, and of identifying those with special weaknesses of personality and temperament that would either unfit them for any type of military service or limit their usefulness to certain less strenuous forms of duty, the army psychologists were also confronted with many special problems which frequently called for new and highly original methods of study. The number and variety of physical injuries required an unprecedentedly large program of vocational counseling for the handicapped. This has meant the re-examination of the merits of tests already standardized and has called for others better adapted to the special conditions for which they are to be used. It has also called for new studies of job requirements, made with special reference to abilities

and skills *not* particularly required in a given occupation and which therefore bring the task within the limitations of men with certain handicaps.

The study of organic brain damage has been given tremendous impetus from the large number of men whose heads were wounded in combat, including those suffering from shell shock. Many tests have been devised for ascertaining the extent and type of damage to the intellectual functions resulting from these injuries, and for studying the more specialized manifestations that appear in some cases, such as motor aphasia or epileptiform seizures, as well as various types of personal-social disturbances. The considerable literature on this topic built up before the war has been ably reviewed by Hunt and Cofer³ (1944). Thus far, few results of the wartime studies in this area have been reported, but the great amount of work that is being carried on should result in much improvement in respect both to diagnosis and to therapy.

THE USE OF TESTS IN THE ARMY AIR FORCES

In perhaps no other branch of the service was the need for wise selection of the men more urgently felt than in air arms of the Army and the Navy. Lack of skill on the part of a pilot or an attack of "nerves" at a critical moment is likely to mean a tragic loss of valuable lives together with the plane, and in some instances it might mean the failure of an important mission as well. The work of the navigator and the bombardier also calls for special qualifications. Gunners likewise must possess certain special abilities, particularly those involving perceptual speed and judgment. Ground crews need to be quick and capable, and mechanics must have the aptitude and skill needed to keep all equipment in order. Failure at any of these points can wreck the entire organization.

In their attempt to meet the tremendous job of selecting and classifying men for these services, the Army Air Forces developed an elaborate system of tests and measurements, of which Guilford and Lacey (1947) have given an account in some 800 pages of description, statistical findings, and evaluation. The book thus provides important source material for the clinical and industrial psychologist as well as for the psychologist in military service, since many of the tests described are adaptable for civilian as well as for military use. Reports of trials of many standard tests such as the Humm-Wadsworth Temperament Scale, the Bernreuter Personality Inventory, and the Minnesota Multiphasic Personality Inventory are also given. The statistical procedures used are described,

³ Chapter 32 in *Personality and the behavior disorders* (J. McV. Hunt, editor).

together with derivation of formulas. In general, these do not differ greatly from those in common use except for the development of short cuts applicable to the employment of IBM machines for scoring, tabulating, and computation.

Factor analyses were used extensively. Correlational matrixes and factor loadings are given for most groups of tests—memory tests, information tests, mechanical ability tests, and so on. The Appendix presents intercorrelations for sixty-five selected tests of various types of ability.

In the examination of a particular test, the air force psychologists were concerned with a number of factors. Internal consistency of the items was for the most part judged by the Phi-coefficient, obtained by finding the relationship between passing or failing an item and belonging to the top or bottom 27 per cent of the group as determined by the total score on the test. The validity of a test was determined empirically in terms of its discriminative value in selecting candidates for the various types of air crew jobs who subsequently made good in their training and in identifying those who failed. Factor loadings were determined for tests that proved worth retaining.

In his presidential address before the Western Psychological Association (1947) Guilford pointed out some of the general implications of the work of the aviation psychologists. He noted how the application of some of the very simple principles learned in elementary psychology sometimes brought about notable improvement in the methods used for training men in various types of skills. For example, recognition of the importance of an immediate knowledge of the extent and direction of errors and the development of devices for providing the men with this information brought about tremendous improvement in the training of flexible gunners.⁴ Many other psychological principles were applied to problems of military techniques with equally advantageous results.

A point which Guilford especially stresses is that low correlations, if based on enough cases to be dependably established, should not be disregarded. He is of the opinion that human behavior is too complicated to be accounted for in terms of a few highly weighted variables. Rather it is a composite of many things, each of which plays a small but important part in determining the uniqueness of each individual. High correlations are thus the exception rather than the rule, but by appropriate combination of many measures, each with its proper factor loading, an aggregate may be reached which should prove both stable and valid for the purpose for which it was designed. The large numbers of men in the army made it possible to use correlations of very low

⁴ A flexible gunner is a man who operates a machine gun in a bombing plane.

magnitude with the assurance that the obtained figure was unlikely to be a chance departure from zero. In civilian life, however, it is not always possible to secure samples of sufficient size to justify the assumption that a correlation as low as those used by Guilford, which sometimes did not exceed .10, represents anything more than an accident of sampling.

As a result of the use of factorial methods, a list of twenty-seven factors emerged. This is many more than has usually been thought adequate for the measurement of human abilities, but, even so, Guilford does not regard it as wholly complete. He notes that twenty of the twenty-seven were found to be important elements for predicting success according to the pilot-training criterion used, but that these twenty factors could account for only about 70 per cent of the variance in that criterion.

Other points brought out by the work of the aviation psychologists may be summarized as follows. No evidence was found for a single factor analogous to Spearman's *g*. Mechanical ability can be defined in terms of mechanical skill or knowledge already acquired, but no evidence for a single factor of mechanical *aptitude* was discovered. A similar statement can be made with respect to clerical aptitude. Only occasional slight indications of possible relationship between the various personality inventories that were tried and success in training were found; in general this type of approach did not appear promising as a means of selecting men for any of the major positions in the air crew. Neither the Rorschach nor the Thematic Apperception Test⁵ was found useful for this purpose.

All in all, it is apparent that more progress was made toward the development of tests of special abilities and aptitudes important for success in aviation than in the determination and measurement of the personal-social factors involved. It is possible that our whole approach to the problem has been wrongly directed, that we have concerned ourselves too exclusively with superficial appearances which can, for the most part, be assumed at will, or which have been derived in large part through the influence of the social milieu. Perhaps we have given too little attention to the bodily accompaniments of emotional changes which are not subject to the control of the individual. Guilford reports

⁵ The Rorschach test was given both in the standard individual form and by two methods of group administration. The individual tests were administered and scored by trained examiners, most of them members of the Rorschach Institute. The TAT was given to men in small groups who wrote out their stories. These were scored in the customary manner.

that the procedures used by M. A. Wenger in the studies of autonomic balance, upon which the latter is now working, showed considerable promise for the study of the psychoneuroses, combat fatigue, and other conditions in which nervous and glandular functions are involved. The suggestion that more attention be paid to physiological syndromes rather than to single physiological symptoms when studying the bodily correlates of temperamental traits is an important one which merits further attention.

PSYCHOLOGICAL TESTING IN OTHER BRANCHES OF THE SERVICE

The work of the psychologists in the air forces provides a sufficient indication of that done in other branches of the service to make it unnecessary to go into the latter reports in detail. Stuit and his colleagues (1948) have presented a detailed report of the work done by the Bureau of Personnel of the United States Navy, while Bray (1948) has given a more general account of the work of the Applied Psychology Panel of the National Defense Research Committee. Many special reports of particular investigations are to be found in the sections on "Psychology and the war" which were published in the *Psychological Bulletin* during the years 1941-1945 and later in the *American Psychologist*.

Some differences between the studies reported by members of the different branches of the service appear. The reason for these differences is not always apparent. For example, Page (1945) found that both the Bernreuter (B1-N Scale) and the Psychosomatic Inventory devised by McFarland and Seitz (1938) revealed highly significant differences in the scores made by diagnosed psychoneurotic patients and those made by a matched group of undiagnosed soldiers chosen from the ranks of trainees. Guilford and his associates rejected the Bernreuter scale for use in pilot selection on the basis of low internal consistency of the items as shown by the Phi-criterion. It is quite possible that these authors placed too much weight on this factor. On the other hand, few other investigators have been able to secure such remarkably high correlations between the scores on these and other personality inventories and psychiatric diagnoses of mental disturbance or other acceptable criteria as are reported by Page. (Ellis, 1946.) On the whole, however, the various groups of military psychologists appear to be pretty well in agreement in respect to the effectiveness for the classification of men of tests grouped roughly under the following heads: (1) achievement tests

dealing with skills and knowledge already possessed,⁶ (2) aptitude tests designed to predict facility in the acquisition of such skills in advance of training, and (3) tests of personality traits and temperamental characteristics that are related to military success. In general, greatest success was attained in the development of measures of the first kind, and these tests were on the whole more useful as predictors of the ability to acquire further proficiency than were measures in which the novice had as good a chance of succeeding as the expert. The difficulty of developing really valid measures of personality and temperament experienced by nonmilitary psychologists was shared by those in the armed forces. Although the reported success along these lines differs to some extent from one military unit to another, it seems not unfair to say that rather less progress was made in this area than in other types of psychological measurement. No really new methods were devised, nor were major improvements made in those available at the beginning of the war. Even on the statistical side, little was done to improve upon the traditional methods of test construction and evaluation.

A COMPARISON OF MILITARY PSYCHOLOGY IN 1941-1945 WITH THAT OF 1917-1918

During the interval between the two world wars, tremendous strides were made in the techniques of mental measurement. The psychologists of 1941-1945 were able to enter upon their task with a far greater body of available methods and with much more advanced technical skill and knowledge than were their predecessors a quarter of a century before. Starting as they did at a relatively late period in the development of the testing movement, their apparent progress was less startling, for it is as true of the growth of a scientific discipline as of any other growth curve that its form is usually negatively accelerated. The work of the psychologists in World War I therefore appears to be well-nigh revolutionary in character, while that of the groups who carried on the psychological work of World War II is clearly evolutionary, involving no great changes in testing procedures and no wholly new statistical methods. Nevertheless, the steady improvement manifested from the beginning to the end of hostilities in the classification of men for specialized military services and in the methods employed for training them, as well as in other areas of psychological service, is evidence that considerable progress was made. Techniques already known were put to

⁶ The General Classification Test is included under this head because the restricted number of types of items included, as well as the nature of these items, is such as hardly to warrant looking upon it as a measure of general intelligence, though it undoubtedly is related to such a measure.

more effective use, short cuts were worked out by which procedures not originally feasible for large-scale work were made suitable for such purposes. Old tests were adapted to the new requirements of military use. In the process of making these modifications much was learned about technical problems of scale construction which was put to use in the development of new tests. All this should prove to be of great value for the future of mental measurement in the postwar world.

Testing and Scientific Investigation

THE MENTAL TEST AS AN INTERDISCIPLINARY TOOL OF RESEARCH

The line that separates one of the social sciences from another has never been drawn very fine. The use of mental measurement as a tool of research more or less common to all has welded another link in the chain which binds these sciences together, for the basic concern of all the social sciences is the analysis and prediction of the behavior of man, considered either as an individual or as a social entelechy. All have to do with the discovery and formulation of rules and principles by means of which unwieldy description may be reduced to more compact form. This, in effect, is the aim of mental measurement.

The biological sciences, too, have been brought into closer alliance with each other and with the social sciences by means of the mental test. That psychology, as the study of human behavior and its development, occupies a position midway between the biological and the social sciences has long been recognized. The study of animal behavior is both a zoological and a psychological problem, and the methods brought to bear upon its solution are essentially those of the mental test. There are tests of animal intelligence and of animal personality. Studies of animal learning have undoubtedly contributed much to our understanding of the principles of human learning. The devices and methods of measurement used in the animal studies are for the most part either similar to or identical with those used for the study of human learning or in tests of human intelligence. The maze, which has a history that extends back into Greek mythology, is still more widely used than any other device for the study of animal learning and intelligence. It was adapted by Porteus (1924)¹ for use as a combined measure of child intelligence and

¹ As used by Porteus, the mazes were printed in the form of a paper-and-pencil test, in which the paths to be followed were traced with a pencil. No overt trial and error was permitted, but at each attempted entry into a cul-de-sac the subject was stopped and allowed one second attempt. Inasmuch as the entire pattern of the maze

personality. More recently, Arthur has included it as one of the series used in her Point Scale of Performance Tests.

The relation of structure to function is a problem of joint concern to the physiologist or the neurologist and to the psychologist. Here, too, animal experimentation has been relied upon in areas where the use of human subjects is forbidden. Studies of the behavior of the intact normal animal have been compared with those made after the infliction of organic brain damage, and these, in turn, with that of human beings in whom such damage has resulted from accident or disease. A study of the qualitative differences in the behavior of normal and abnormal subjects has made it possible to develop a system of signs from which the existence of brain damage can be inferred with a fairly high degree of probability in cases where the occurrence of such injury was unknown or unrecorded. More will be said about these signs in the following chapter.

Answer to the old question of the relationship between the phylogenetic and the ontogenetic sequences of behavior is no longer sought in terms of casual observation or appeal to authority, but by means of actual measurement. Identical problems² are set for different species, including man, and for different age levels within each species. Both the qualitative and the quantitative aspects of their performances are compared. Hunter's study of the delayed reaction in man and animal (1913) is one of the earliest in which an actual "mental test," similar for all subjects, was used to measure the differences in performance of subjects covering a fairly long stretch of the phylogenetic sequence. In the more intensive studies of a child and a chimpanzee reared together for a period of nine months which covered, roughly speaking, the latter half

could be seen, it was Porteus's belief that the cautious person would be likely to stop and look ahead at each turning and so avoid at least the more immediate and obvious errors. The impulsive person, on the other hand, would tend to dash along as rapidly as possible without pausing to examine where his path might lead.

As far as the limited situation provided by this system of mazes is concerned, there is justification for the conclusion that whatever residual is left, after differences that can be accounted for on the basis of the experimental error of measurement and variations in intelligence have been equalized, is in all probability due to a temperamental factor similar to that described by Porteus. Whether or not an individual's performance on the maze is an adequate sample of his behavior in other situations is another matter. While such evidence as is provided by ratings of associates and by more or less well-controlled observation of behavior suggests that the maze test has some validity as a measure of the caution-impulsiveness continuum among persons of similar age and intelligence, the correlation is not high enough to warrant more than very tentative conclusions without other data by which they can be supported.

² Some adaptations in the actual setup of the tasks may occasionally be necessitated in order to allow for differences in the size and structure of the subjects to be compared, but the essential features of the problems to be solved are kept the same for all.

of the period of infancy, the Kelloggs (1933) made very extensive use of mental tests, including both the formal series developed by Gesell, and others especially devised for use in their experiment. Not only intelligence tests but motor tests of various kinds were used. This study derives its value quite as much from the fact that uniform measures were used for both the human and the animal infant as from the fact that both were kept under like conditions of environment and training during the experimental period, for the effect of the latter could not have been adequately determined without the use of the former.

Through the application of measurement to animal behavior, many scientific problems which could not otherwise be readily subjected to experimental investigation have been approached. The work of Tryon (1934) on the inheritance of maze-running ability in rats or the allied studies by Rundquist (1933) on the inheritance of spontaneous activity are examples. Because human generations are so widely separated in time, and also because the methods of selective mating employed by both Tryon and Rundquist cannot be used with human beings, the question of the inheritance of mental ability as well as of physical characteristics in man has long been a matter of controversy. The fact that children are ordinarily reared by their own parents makes it difficult to separate the effects of environmental stimulation from those of biological inheritance. To the extent that the basic laws of inheritance are similar for man and animal, the studies just mentioned appear to demonstrate beyond reasonable doubt that variances in behavior as well as in gross structure³ may be passed on from parent to offspring, even when environmental conditions remain the same.

Sociology and social psychology are so closely related that it is not always easy to say where one begins and the other leaves off. Theoretically, the former is chiefly concerned with the relationship and interaction of groups and with defining and describing the characteristic features of different cultures. Sociology also deals with the origin and effect of social movements and with the social changes brought about by alterations in the political or economic structure of a society. In contrast, social psychology deals mainly with the social interaction of individuals or of small groups, and with the factors by which such behavior is modified. Anthropology, although it is defined as the science which deals with the development and history of man in relation to his environment, has actually been but slightly concerned with modern civilized

³ This is not to say that the behavioral variations in question are unrelated to structural differences. That they are based upon changes in the structure or organization of the nervous system can hardly be doubted, although the exact nature of these changes is unknown to us.

societies, centering its attention upon the more primitive groups, both ancient and modern.⁴

In all these fields, modern research workers are coming to see that mental tests, in the broader interpretation of the word, not only provide much valuable information in their own right but may also serve to reveal possible sources of error arising from unsuspected inequalities of sampling or from confusion between superficial appearances and the more basic factors underlying them. For example, if two primitive groups are found to differ markedly with respect to the complexity of their language structure or of their social organization, the basis for the difference may be sought either in the conditions under which each has lived or in a more fundamental difference in mental capacity. The use of mental measurement in sociology and anthropology has not only helped to answer new questions as they arise but has raised a good many questions as to the soundness of traditional beliefs based upon investigations in which the control of conditions was inadequate or the descriptive reports were subject to more than one explanation or interpretation.

That psychology and its allied field of abnormal psychology have profited greatly by the use of tests and testing procedures is too well known to require more than very brief mention. Among the areas in which the use of tests has proved most valuable are the following: (1) the studying of the intellectual abilities and personality characteristics of the physically and mentally handicapped; (2) the identification, classification, and etiological diagnosis of personal-social difficulties among persons who, although classed as "normal," are nevertheless so hampered by these problems that they are unable to function at the level of which they are potentially capable; (3) the diagnostic study of the mentally abnormal; (4) the development of therapeutic methods and the measurement of progress under treatment; (5) vocational guidance and industrial classification; and (6) the development of prognostic measures.

In the greater number of studies in the field of general psychology that are being carried out at the present time, some use of tests is reported, even though the topic of investigation may have little to do with the question of individual differences. Inasmuch as many of these investigations involve the determination of the effect of some interposed condition upon the behavior of groups, a comparison between an experi-

⁴ This statement should be taken in a relative rather than in an absolute sense. Anthropologists have made many studies of the characteristics of modern man, both for their intrinsic interest and also in order to provide a basis for comparison with his more primitive forebears or contemporaries.

mental and a control group is usually required. Unless these groups are matched for initial ability, such comparisons are apt to be misleading. In many instances, the methods of study used are essentially those of the mental test even though they may not be so designated. In the more specialized area of educational psychology the use of tests is even more nearly universal, in part because of tradition, since the testing movement was primarily developed to meet educational needs, and in part because educational problems continue to demand their use.⁵

The use of a common tool of research serves to emphasize the essential unity of science. Always, scientific progress has hinged upon measurement and the study of the relationships among measured phenomena. Until such measurement is made possible, no exact formulation of these relationships can be worked out, for while grossly quantitative ideas can be expressed in such descriptive terms as "more" or "less," these expressions are not sufficiently refined, nor are they sufficiently divorced from the concrete situations to which they refer, to make them suitable for the expression of general rules and principles.

Mental measurement, as the last few chapters of this book have endeavored to show, has now advanced far beyond the range which it was originally designed to cover. No longer is a mental test necessarily a test of intelligence. No longer does the work of the clinical psychologist stop short with the administration of a Binet test interpreted by rule of thumb, with perhaps some half-dozen performance tests of uncertain significance thrown in for good measure. No longer can the prospective mental examiner complete his training by means of a six weeks' course in summer school. The area now covered by the methods of the mental test comprises practically every aspect of human behavior and its aberrations. It is true that the soundness of many of the methods used has not been sufficiently proven, that the experimental errors of the tests are frequently so large as to render them untrustworthy for any but very crude uses, that excessive claims have been made for instruments later proved to be of little worth, and that relatively few (if any) of the tests now available have been reduced to a state of mechanical accuracy that renders them approximately as effective when used by the tyro as by

⁵ A rough classification of the forty-eight articles included in the 1946 volume of the *Journal of Educational Psychology* shows that only four are unrelated to the field of testing. Of the remainder, nineteen have to do with standard tests and their application to specific educational or psychological problems. Eight deal with various aspects of statistical methods closely related to the field of tests and measurements, such as factor analysis, item analysis, and the like. Informal methods similar to those of the standardized test, including testing devices especially developed for use in a particular investigation, are reported in nine studies, while in six others tests are used chiefly as tools for describing the subjects or the conditions of investigation. The two remaining articles are theoretical discussions of tests and testing.

the expert.⁶ In spite of all these shortcomings a substantial advance has been made toward the quantitative expression of facts which only a short generation ago would have been thought by all except a farseeing few to be unsusceptible of objective measurement.

The methods and concepts of the mental test have now penetrated into most of the social and biological sciences. As a result, new questions have arisen and old questions have found new answers. For the course of civilization is determined by human relationships. Only as we are able to analyze these relationships into their basic patterns shall we become able to harmonize the design, see where our errors have been made in the past, and prevent their occurrence in the future. While the magnitude of this task puts its solution well beyond the scope of any single method of approach, its importance is so great as to demand that no promising mode of attack be overlooked. That the mental test as developed and expanded during recent years provides one such method is a widely accepted belief for which an important body of supporting evidence is accumulating.

⁶ A few of the tests of educational achievement approach this level fairly closely if they are regarded as indicators of what the subjects are now able to do and not of what they are potentially capable of doing in respect to the particular skill measured by the test. It is questionable, however, whether any instrument designed for the measurement of human ability or the prediction of human conduct can ever be devised which will be as nearly free from the necessity for special training on the part of the user as is the common household gadget. Among other things, there is always the question of detecting the cases in which the instrument has led to misleading conclusions. As a rule, when something goes wrong with a mechanical gadget, it advertises the fact by simply refusing to work or by giving absurd results. In most cases, too, when a mental test that has generally been found useful is rendered unsuitable for use in a particular case by reason of special coaching or other unusual conditions, the experienced and well-trained examiner can detect signs that something is wrong. These signs, however, are so subtle that they are unlikely to be noted by the amateur.

The Use of Tests in the Study of Group Differences

THE SCIENTIFIC VALUE OF COMPARATIVE STUDIES

Scientists in many fields are finding that a comparison of the mental and behavioral differences among groups separated on the basis of some peculiarity that can be objectively measured or categorized frequently provides an acceptable substitute for direct experiments that they are unable to make. Innate characteristics such as race or sex cannot be changed at the will of the investigator. Social and ethical considerations also set limits to his experiments. The social psychologist who wishes to study the relation of religious beliefs to other aspects of behavior cannot rear one group of infants in the Jewish faith, a second as Roman Catholics, a third as Protestants, a fourth as Buddhists, and so on, in order to test whatever hypotheses he may have formed. Nor can the one who is interested in the mental effects of venereal disease deliberately subject his experimental subjects to syphilitic infection merely for the sake of scientific study.¹ Nevertheless, although such drastic measures would ordinarily not be considered, advantage can be taken of those conditions which already exist. Although we cannot alter the sex of any individual at will, we can study the characteristics of persons whose sex differs. We cannot compel religious beliefs, but with the imposition of the necessary controls we can investigate the differences in the behavior and attitudes of groups whose members belong to various religious denominations. We would not intentionally inflict damage upon any human being, but we may compare the characteristics of those who have been the victims of accident with the corresponding traits of their mates who are physically normal. Although studies of this kind are always and inevitably

¹ In the control of certain diseases that have taken a heavy toll of human life in the past, such as yellow fever, which was experimentally proved by the American Army Board under Walter Reed to be communicated through the mosquito, just such heroic methods as these have been used. Only in grave situations, however, would they be tolerated.

subject to possible inequalities of sampling, particularly when the available numbers are small, if careful attention is given to matching when a comparison is to be made between an experimental and a control group,² or if adequate normative data are available for use when a selected group is to be compared with the generality, serious errors of this kind can usually be avoided.

Comparison of the mental and social characteristics of contrasted groups thus holds almost endless possibilities for scientific investigation. With proper control of conditions, the method permits the testing of hypotheses by analysis of the variance into its component factors, by correlational analysis, and by other statistical methods for determining the amount of confidence that can be placed in such differences as are found, or in the theories that have been proposed to account for these differences. Only a few examples of these studies can be mentioned here, but these may serve to illustrate some of the trends of modern thought and the manner in which data of this kind are utilized in the investigation of scientific theories and popular beliefs.

EXTREME DEVIATES IN INTELLECTUAL ABILITY

The most extensive and thoroughgoing studies of the development of intellectually gifted children are unquestionably those directed by Professor Lewis M. Terman of Stanford University. The four large volumes of his report, including that by Cox on the childhood characteristics of men of genius (1926), cover a quarter of a century—from 1921, when the study was begun,³ to 1945, when the latest follow-up investigation was completed.

One of the main contributions of the 1925 report was that of exploding many of the popular myths concerning the physical and personal characteristics of bright children. Instead of being small and weakly, they

² It may be well to point out, since the point has so often been overlooked, that when the method of matched groups is employed, if tests of significance are used, allowance must be made for the correlation between members of the groups which is thereby introduced.

³ For more than a hundred of the subjects, the investigation covers even a longer period than that indicated above. Terman's interest in exceptionally bright children dates from the time of his doctoral dissertation at Clark University in 1905 (published 1906). From the time that he began work on the standardization of the 1916 revision of the Stanford-Binet, he gave special attention to the cases whose performance on the test was unusually high. In *The measurement of intelligence* (1916) several pages are devoted to this subject, and brief case studies of ten such children are presented. In *The intelligence of school children* (1919) two entire chapters are given up to a discussion of superior mental ability in childhood. Many of the cases described in these two earlier volumes as well as others located at about the same time were carried over for inclusion in the main study begun in 1921.

were found to be above the generality of children in size, health, and physical vigor. Although there were wide individual variations in personal-social traits, the gifted children, on the whole, were more popular than most children of their age. They were modest rather than conceited, outgoing and socially inclined rather than introverted and withdrawn. They were unusually free from the personal eccentricities so often attributed to the highly gifted individual.

After disposing of these misconceptions, and analyzing the biological and environmental background from which these children came, Terman was confronted with two major problems. First and most obvious was the question of the relationship between childhood brilliance of the kind manifested by extraordinary scores on intelligence tests and adult accomplishment. This is in part answered by the follow-up study reported in 1947, though more complete information will be obtained with the progress of time.⁴ Some indirect evidence on this head is also provided by Cox's study on the childhood of men of genius, in which a rather striking resemblance is reported between the early accomplishments, interests, and general behavior of the three hundred famous men whose biographies she studied, and that of the children in Terman's group. Although none of the latter have as yet manifested accomplishments that are at all comparable with the geniuses of history, it is perhaps unfair to expect that this would be the case. In addition to the age factor previously mentioned, the samples are by no means comparable as to rigidity of selection. Cox attempted to choose the three hundred men of greatest eminence who were born during the four-century period from 1450 to 1849, for whom apparently dependable biographical records written in French, German, or English were available. A further criterion was that the fame of these men should rest upon accomplishment and not upon such accidental conditions as royal birth or other factors not of their own making. Fourteen different nationalities are represented. In contrast, the children of Terman's group were born within a total span of twenty-five years, with well over 80 per cent born within a ten-year period. With a few exceptions, all were residing in one or another of five California cities at the time of selection. Practically all of the gifted children and about 75 per cent of their parents were American-born.

On the whole, and in spite of the apparent absence of top-flight geniuses in Terman's group, the question as to whether there is a relationship between mental-test performance in childhood and adult accom-

⁴ At the time of the 1945 follow-up, the mean age of the subjects was slightly under thirty-five years. Very few people, especially those in the professional and managerial classes, have reached their maximum level of accomplishment at so early an age.

plishment must be answered in the affirmative. In 1940, when their average age was slightly below thirty years, more than 70 per cent of the males were engaged either in the learned professions or in semi-professional or higher business occupations. This proportion is more than five times that reported for unselected males of corresponding age and residence. Their incomes were approximately 75 per cent higher than the average earned by employed persons of corresponding age and sex at the same period, and were somewhat higher than those of the average college graduate, even when no allowance is made for the fact that a third of the gifted subjects had not graduated from college.

More important from the standpoint of creative genius are the scientific, literary, and artistic productions of this group. Almost four hundred of them have had material published, though only forty-one are professional writers. By far the greater number of these publications are reports of scientific research, though poetry, fiction, essays, and practically every other form of literary composition is also listed. Around thirty are professional artists or musicians; others are engaged in some form of dramatic art. That some of these young persons may still find place among the ranks of the truly great is by no means impossible.

This brings us to the second question. What is genius? Can it be defined in terms of the intelligence quotient alone? Is the IQ a sufficiently dependable sign from which later achievement may be predicted? That it is one of the signs is clear from the facts that have just been reported. But is it enough?

Not all of Terman's gifted children have achieved outstanding success. Not all of Cox's famous men are reported to have shown really remarkable ability during childhood, though it is of course possible that lack of recorded information is responsible here. Nevertheless it is apparent that there is not a one-to-one relationship between intelligence test score in childhood—or even in youth—and accomplishment in later life. The correlation is positive and fairly high but it is by no means perfect. How can we account for the differential?

This is the major question that Terman set himself to investigate in the 1940-1945 follow-up of his group of gifted children. After making an exceedingly detailed and careful study of their accomplishments, behavior, and personal characteristics as young adults, three judges went over the records of all the men in the group⁵ and classified them into three groups on the basis of their apparent accomplishment up to that

⁵ Because such a large proportion of the women had neither desired nor entered upon any career other than marriage and parenthood, for which no objective evidence of success is available, it was not deemed possible to include them in this part of the study.

time. A comparison was then made of the background and personal characteristics of the 150 men in the top group (designated as A) and of the 150 in the bottom or C group. Inasmuch as all were at the 99th percentile or better in tested intelligence it was felt that this procedure might throw light on some of the nonintellectual factors making for eminence.

Although the judges who made the classification had been careful not to take tested intelligence into account, a slight difference in favor of the A group was nevertheless found, even in the childhood records. More significant is the fact that with advancing age the two groups tended to draw farther and farther apart both in respect to tested intelligence and in scholastic success.⁶ This suggests a possible difference between the two groups in age at intellectual maturity. It may be that one of the characteristics of the extremely able person consists in an exceptionally long period of mental growth. Possibly allied to this is the fact that the members of the A group came, on the average, from better intellectual stock. Their parents were more highly educated, and those of their siblings who were tested earned higher IQ's than did those of subjects belonging to the C group. It is also possible that there is a difference in the direction taken by the experimental error of the test in the two instances. In the case of the A group, the original standing may have been slightly too low; for the C group, it may have tended to be too high. Each of these possible explanations would tend to emphasize the actual importance of the intellectual factor beyond that which appears at first sight.

One of the major differences between the gifted children and those of average intelligence that was noted at the time of the original study was their zest in living and doing. They read many more books, had an amazing fund of general information, made more collections, and had many more hobbies. They had a keen desire for knowledge about almost everything with which they came in contact. Teachers ranked them as somewhat more popular than the average child and with distinctly greater capacity for leadership.

There are indications that something analogous to this intense *joie de vivre*, with its drive toward learning and doing, its self-confidence, and its clearly directed interests, is one of the important factors which separates the A group from the C in adult life. The occupational histories of the latter revealed more and longer periods of unemployment and more frequent shifts from one job to another. When questioned as to whether or not they liked their present jobs, 98 per cent of the A's but

⁶ Intelligence tests were given in 1922, 1928, and 1940.

only 75 per cent of the C's answered in the affirmative. More than twice as many of the A's as of the C's stated that they had definitely chosen the line of work in which they were engaged; two thirds of the latter claimed to have just drifted into it. Only 9 per cent of the A's but 43 per cent of the C's said they would choose another kind of work if they had the opportunity. While all this suggests that vocational misplacement may have been a factor of considerable importance in determining the relatively low level of success among the C's, it is also likely, as Terman points out, that some of their expressed dissatisfaction may have been simply a mode of rationalization for their lack of success. Actually it is likely that a circular relationship exists, with unfavorable temperamental traits accentuating poor accomplishment and vice versa.

The improbability that vocational misplacement is alone responsible for the difference between the two groups is indicated by a number of associated factors. There is a considerable difference in their high school records and a very marked difference in their college success. More of the A group graduated from college and very many more went on for higher degrees. By 1940, thirty-three of the A's but only one of the C's had earned the Ph.D. More than half (51.9 per cent) of the A's but only 14.3 per cent of the C's were graduated from college with honors. The actual difference here is much greater than the above figures indicate, since they refer only to those who remained to graduate. If corrected to values based on the total number of cases, the contrast is much greater—46.7 and 4.6 per cent, respectively. Something more than the ascertained difference in intelligence is needed to account for so great a discrepancy in educational accomplishment.

The results of the Strong Test of Vocational Interest provide some further clues. Some of the men in the C group failed to earn a high score on any of the occupational keys. This suggests either that their interests were of so unusual a pattern that they did not correspond to those shown by men in any of the common occupations or that they were not motivated toward any particular line of work. The mean number of occupations for which the interest scores were B⁺ or better was reliably higher for the A's than for the C's. Moreover, the score on Strong's occupational level key, which is generally regarded as an indicator of the level of aspiration, was markedly higher for the A's than for the C's. All this again points to the conclusion that the strength of the interest felt in the work that one has chosen to do and the zest with which he attacks it are important factors in success.

In play as well as in work, the A's showed keener interests and greater versatility. According to their own statements, there were many more recreations that they enjoyed very much. Their personal lives, too,

appeared to be more satisfactory. Fewer of their marriages have terminated in divorce; their scores on the Terman test of marital adjustment were reliably higher, and practically all ratings⁷ of personal-social characteristics and emotional adjustment show some advantage for the A as compared to the C group. All this does not answer the question as to the nature of genius, or the simpler question which has to do with the factors that make for the extraordinary accomplishment by which genius is recognized.⁸ But it does afford some clues.

Undoubtedly, a good deal of confusion exists as a result of the tendency of some people to think of genius in terms of intellectual capacity as indicated by standing on tests of intelligence, while others regard it as practically synonymous with attainment of an exceedingly high order. Because there are many circumstantial factors that may interfere with achievement, the concept of genius as *capacity to achieve* is probably the more valid of the two. But the assumption that this capacity can be measured by means of intelligence tests alone is not borne out by such data as are available. One cannot justifiably assume that all the cases in Terman's group who reached or exceeded the median IQ of approximately 160⁹ reported by Cox for her group of three hundred historical geniuses necessarily possess the capacity for equally outstanding achievement. They may be lacking in other essential characteristics.

Granting that achievement is an imperfect criterion for capacity to achieve, it is nevertheless the most objective of those available at the present time. Any other involves unproven assumptions with respect to the factors underlying achievement. The use of such criteria is therefore tantamount to circular reasoning. But if we compare the available facts regarding the characteristics of those who have achieved greatly and those whose accomplishments are small, we may arrive at some conclusions, tentative though they necessarily are, as to what these factors may be.

That general intelligence occupies a place of prime importance in

⁷ These include self-ratings, ratings by wives, and ratings by field workers on a large number of different traits.

⁸ That "mute, inglorious Miltons" have existed in every age and clime is very probable. Genius and fame are by no means synonymous terms. Nevertheless I suspect that the number of those who truly merit the designation of "genius," but whose names have never been sounded by the trumpet of fame, is smaller than many have supposed. And as modern facilities for communication continually bring us closer to a type of existence analogous to that of the proverbial "goldfish in a bowl," the number of such persons is likely to decrease still further.

I question, moreover, whether the concept of the "mute, inglorious Milton" is either scientific or useful. Certainly a Milton may be inglorious by reason of opportunity and circumstance. But can he be mute? And if he is, by what means are we to know that he is a Milton?

⁹ After an estimated correction for unreliability of estimate and insufficiency of data.

the list seems indisputable. There is no dependable record of an individual who had not at least average intellectual ability attaining first rank in any field of importance for human progress; it is questionable whether such outstanding accomplishment is possible for anyone who is not of superior intelligence.

Both Terman's comparison of his A and C groups and the biographical material assembled by Cox for her three hundred historical geniuses suggest that a factor of hardly less importance than that of intelligence is something analogous to the volitional factor which Spearman designated as *w*, or "purposive consistency." In summarizing the comparisons between his A and C groups, Terman notes that there is nothing in which the two present a greater contrast than in the drive to achieve and in all-round social adjustment. The drive to achieve, moreover, is selective, not random; and it is realistic, not rooted in fantasy. It is joyous achievement in which obstacles do not deter but spur to greater effort.

Whether we are dealing here with a factor which is at least in part innately determined or whether its variance can be wholly attributed to differences in experience and early training is not known.¹⁰ No valid tests or other methods of measuring it have been developed, but the biographical data point irresistibly to its existence and to its importance for achievement. It would seem that the trait—if it may be called a trait—should be susceptible to objective study and that the term used by Spearman, viz., "purposive consistency," describes it as well as can be done at the present time.

A number of aspects of general social adjustment on which data are reported for the two groups show fairly reliable differences in favor of the A group in 1940, but since the facts are not given it is impossible to say whether or not a similar difference was manifested in childhood. To what extent the superiority of the members of the A group in social traits was a reflection of their greater success along other lines and to what extent it was a factor in the production of success cannot be stated

¹⁰ Some evidence on this head, however, is provided by the consistency of the difference between the A and the C groups in respect to this characteristic over a period of eighteen years. In 1922, when the subjects were still children, they were rated by parents and teachers on a series of twenty-five traits. Among these were four which, as defined, are at least partially related to the characteristic we are discussing. These were (1) prudence and forethought, (2) self-confidence, (3) will power and perseverance, and (4) desire to excel. The critical ratio of the differences in the combined ratings on these four traits as given by parents and teachers was 3.96. Six years later, in 1928, a similar method was employed but only two of the four traits were rated. These were (1) perseverance and (2) desire to excel. Again the A group reliably exceeded the C's (critical ratio, 5.23). While this is by no means incontrovertible evidence that the trait is innately determined, it at least shows that by the age of six to twelve years, individual differences have become sufficiently well established to make it improbable that variations in later experience will greatly affect their pattern.

with assurance for lack of information regarding the sequence of events. As in so many other instances, a circular relationship may exist.

On the environmental side, it is apparent that the A group had a considerable advantage over the C's, and that Cox's famous men also came, for the most part, from homes of better than average standing with respect to material and cultural surroundings. They had a better start. Also, the frequency of divorce or separation of the parents in Terman's group was somewhat greater among the C's than among the A's, but a fair number of broken homes are recorded for both groups. Among those of the subjects who had married before 1940, the differential rate of divorce is even greater than among their parents. In his study of marital happiness previously mentioned, Terman likewise found that divorce is more common among the children of divorced parents than among the generality. It is impossible to say whether this is mainly attributable to biological or to social inheritance.

In the homes from which these subjects came and to a far greater extent in those which they later established for themselves a marked difference in the material and cultural advantages is to be noted. Although to some extent the superior character of the conditions under which the members of the A group are living may be due to fortunate accident and to the initial advantages provided by their families, the overlapping between the groups in the latter respect is too great to render the hypothesis tenable as a factor of great importance in producing the difference.

That sheer good fortune may have been a factor in the success of some is in all probability true, and the method of selecting the groups would undoubtedly tend to place the lucky ones in Group A and those upon whom fate has frowned in Group C. But the blind goddess is less fickle than she seems. Opportunities are not random events, falling like rain upon the just and the unjust. To a far greater extent than many realize, opportunities are self-made. And this raises the question. Is the ability to create opportunities, to recognize and take advantage of the fluctuations of chance a special talent in itself; a talent that is allied to but not identical with the kind of ability that we call "general intelligence"? Such talent, if it exists, would separate the realist from the dreamer, the man of foresight from the one who thinks only of the present. It involves the ability to organize experience in terms of some foreseen goal. It enables its possessor to integrate his experience in such a way that events and circumstances are no longer viewed in isolation but in the light of their possible contribution to the attainment of that goal.

The question of the factors that make for the extraordinary accomplishments which we designate as "works of genius" has been taken as an

example of the contributions that the study of special groups can make to scientific knowledge. The examination of extreme cases and of the factors that give rise to them is rarely undertaken for the sole purpose of learning more about subjects of that type. In the course of such studies other questions arise that may have equal or even greater importance in comparison to those which formed the starting point of the investigation. Particularly is this likely to be true when contrasted groups are used, as was seen in the brief survey given above of Terman's study of the contrasts between successful and unsuccessful men of outstanding intelligence. We turn now to the opposite end of the scale of intellectual ability—to those whose mental capacity is so slight that they are commonly regarded as mentally defective or, in more popular speech, as "feble-minded."

Even in these comparatively enlightened days it is not uncommon to find persons who adhere to the view that an IQ below 70 always and necessarily connotes "feble-mindedness." They are willing to consign an individual to an institution for life on the basis of an IQ of 69, and are alike unmindful of the test used, the conditions under which it was given, and the experimental error of measurement. That such an attitude cannot be justified in the light of modern science should be obvious to all who have read this book.

Kanner's (1935) use of the term "intellectual inadequacy" to designate a level of ability falling below that required to cope successfully with the requirements of any specified set of external conditions calls attention to the fact that, in determining who are inadequate, account must be taken of the demands likely to be made upon the individual as well as of his own intellectual powers. Other characteristics besides intelligence also play a part in determining the outcome. Here, as well as elsewhere, we are dealing with probabilities and not with certainties. A clear recognition of this principle would do away with many foolish pronouncements from which textbooks in educational psychology or similar fields are by no means free, and would also prevent many unwise diagnoses of individual cases on the part of clinicians whose training in statistical theory and method has been poor.

Figure 40 illustrates the two ways of thinking about the matter. The curve on the left (marked A) indicates the all-or-none view, in which a sharp line is drawn at some point on the distribution (usually IQ 70), with all who stand above that point classed as "normal" and all who fall below it as "feble-minded."

The curve on the right indicates the correct view in terms of probability. As one moves out toward the extreme left of this curve the likelihood that a given individual will prove "adequate" to meet the demands

of a particular situation becomes progressively less.¹¹ This is indicated by the increasing frequency of black dots (representing chances of failure) toward that end of the distribution. At the extreme left, that is, at the very low IQ levels, the likelihood of success under ordinary conditions of life becomes so small that it can be disregarded. Even there, however, if one makes use of the concept of intellectual adequacy, success of a kind is no longer regarded as impossible. Persons of exceedingly low IQ may be intellectually adequate for the requirements of life within the protecting walls of an institution. There some of them may

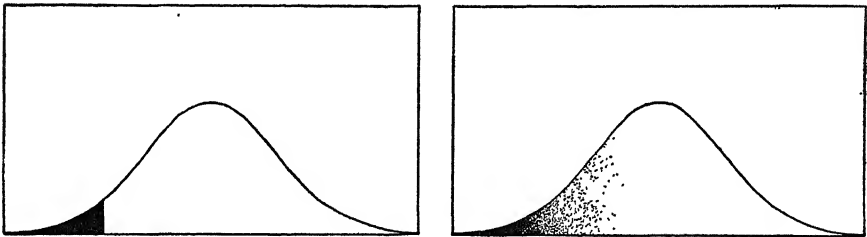


FIG. 40. CORRECT AND INCORRECT REPRESENTATION OF THE DISTRIBUTION OF THE INTELLECTUALLY INADEQUATE IN RELATION TO MENTAL TEST SCORE. (Reproduced by permission of the D. Appleton-Century-Crofts Company from *Developmental Psychology* by Florence L. Goodenough.)

learn to do a certain amount of useful work under supervision, while others of lower mentality may perhaps learn to feed and dress themselves with varying amounts of help.

Since the absolute level of IQ obtained on a particular test is no longer regarded as sufficient evidence of the probability that a given individual will be able to cope successfully with the conditions of life to which he is likely to be subjected,¹² at least two additional types of measure are needed for purposes of prediction. The first will be an esti-

¹¹ One may also think of the curve as representing the proportionate number of individuals at successive IQ levels who will be able to get along successfully under ordinary conditions of social and economic stress, no consideration being given to those who fail for other than intellectual reasons.

¹² Americans who accept the first part of Tredgold's definition of mental deficiency as "a condition in which the afflicted individual, owing to imperfect or incomplete cerebral development existing from birth or from a very early age is unable to perform his duties as a member of society" have generally been loath to admit the final qualifying phrase—"in the position of life to which he is born." Yet even in America, the child of a manual laborer, born in a low-class neighborhood, who attends a school where most of the children are of a social class similar to his own, is not faced with the same intellectual requirements as is the son of a successful physician living in a superior neighborhood. Maller, it will be recalled, found that the *average* IQ of children attending schools in different sections of New York City ranged all the way from 76 to 120.

mate based upon measurement, rather than a direct measure of the thing itself. Its aim is to make the best possible estimate of the intellectual standards of the environment with which a given individual will be most likely to have to cope as an adult. This involves finding the regression weight for some measure of adult environment upon child environment,¹³ when the measure chosen is the best available sign of the intellectual requirements of a given environment. Although the experimental error of such an estimate would be fairly high when made during the childhood of the subject under consideration, it should still be possible to derive a formula that would add materially to the value of the IQ as a means of predicting the future intellectual adequacy of a given individual, since it would take account of the most probable future demands that will be made upon him and would thus help to answer the question: Adequate for what?

The second question is this: Since qualitative as well as directly quantitative differences in intellectual ability undoubtedly exist, can any special signs be noted in the course of a psychometric examination that will help in the practical classification of individuals as good or poor risks from the standpoint of their intellectual adequacy for meeting the socioeconomic demands that they are most likely to encounter? If such signs exist they would be of great value, particularly in the cases of persons of borderline intelligence, for whom the chances of success and failure appear about equally balanced.

As is so often the case, we turn first to Binet as a possible source of ideas. Binet's many and detailed descriptions of the nature of intelligence were never crystallized into a formal definition, but Hollingworth (1920) has summarized his views of its essential attributes under three general heads: (1) the ability to take and maintain a given mental set, (2) the capacity to make adaptations for the purpose of attaining a given end, and (3) the power of autocriticism. It is interesting to compare this analysis with Terman's findings on the differences between the most and the least successful members of his group of high-testing children and in so doing to speculate on the possibility that our present intelligence tests are not as discriminating as they should be with respect to the first two of the three attributes that Binet believed to be of chief importance. Certainly it is along these lines that the two groups studied by Terman mainly differed. The possible advantage of supplementing the intelligence tests now available by others particularly designed to measure

¹³ Assuming, of course, that we are attempting to predict the intellectual adequacy of an adult on the basis of his childhood IQ. In the case of a person tested after he has already arrived at maturity, an estimate of future probabilities is no longer needed, for one can then consider the existing facts.

these two capacities should not be overlooked by those interested in test construction.

The fact that intellectually retarded individuals are markedly lacking in the three characteristics mentioned by Binet was particularly stressed by Hollingworth in the monograph just cited; it has also been noted by most others who have had extensive dealings with them. Many find it hard to persist in any activity for more than a few minutes. Others persevere rather than persist. They repeat some simple act over

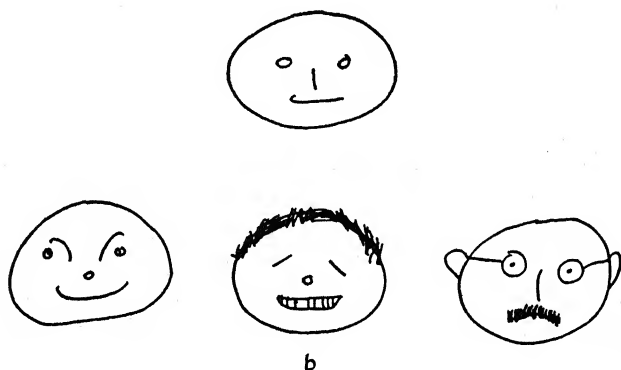


FIG. 41. BRIGHT CHILDREN MAINTAIN THEIR INTEREST IN A MONOTONOUS TASK BY VARYING THEIR MANNER OF PERFORMING IT. (Reproduced by permission of the D. Appleton-Century-Crofts Company from *Developmental Psychology* by Florence L. Goodenough.)

and over without essential change, but their attention is directed toward the act rather than toward any more remote purpose to which the act may contribute. Lewin (1935) has given a neat example of this by means of an experiment in which two groups of children, one of normal intelligence, the other feeble-minded, were told to draw "moon-faces" similar to those in Figure 41a. They were urged to keep on making these faces as long as they were willing to do so. Both groups soon became bored with the task, but the bright children found ways of enlivening it, such as are shown in Figure 41b. The feeble-minded, whose mental set—if we may call it such—was directed only toward the activity of making moon-faces like those of the pattern with no real comprehension of the more abstract idea of "a face," had no such ways of making adaptations in their manner of achieving that end. They kept on drawing without essential change until the monotony of the task overcame them. Then they stopped.

This inability to make adaptive changes in the pattern of a habit once established led Lewin to formulate his concept of *mental rigidity*

as the basic characteristic that separates the mentally defective person from those of normal intelligence. Goodenough (1931) has shown how the tendency to perseveration of habit, with its associated inability to "make adaptations for the sake of achieving a given end," affects the drawings of feeble-minded children, making it very difficult for them to add new features to their art products without disturbing the relationship of the parts previously drawn. For example, in the drawings of very young children the trunk is usually omitted and the legs are attached to the head. When the trunk is added to the scheme, normal children usually experience no difficulty in placing it. Feeble-minded ones, on the contrary, are likely to stick to their old pattern. They draw the head and attach the legs to it as before; then, as there is no other place for the trunk, it is usually suspended between the legs. Even when the legs are not added until after the trunk has been drawn, they are still likely to be attached to the head or sometimes even to the brim of the hat. Similar evidences of this nonadaptability of habit in the child of inferior intelligence are seen when, in drawing the human figure, he attempts to change from the full-face position to the profile. Two sets of features, one appropriate to each position, often appear along with many other bizarre errors. Normal children and even adults also experience difficulty when changing from a well-established habit to a new one, but they make the necessary adaptations much more quickly than the feeble-minded do. Here Binet's third attribute comes into play. The normal person is likely to realize his mistakes and try to analyze them. The feeble-minded are much less likely to see that they have made an error unless the recognition is forced upon them by obvious failure. And in such cases, although they know that they have not succeeded, they are rarely able to see what is wrong.

All this suggests the possibility of utilizing clinical signs obtained in the psychometric examination to supplement the quantitative findings and aid in the diagnosis of doubtful cases. Wechsler has shown that mentally defective persons usually do better on the subtests comprising his nonverbal scale than they do on the verbal scale, and that the score on the arithmetic subtest is likely to be particularly poor. Goodenough (1925) showed that backward children do relatively better on a test of word meaning than on one of paragraph meaning, while the opposite is true in the case of bright children. Burt (1921) was so impressed by the qualitative differences in the drawings of normal and feeble-minded children that he was of the opinion that in most cases a differential diagnosis could be made on this basis alone. Rorschach workers have also pointed out a number of signs by which the intellectual level of the subject can be judged. An examination of the nature of these signs may

provide further information as to the qualitative aspects of intelligence.

The study of extreme deviates in mental ability has thus served to throw considerable light on the nature and organization of mental ability. It has also suggested a number of additional points of attack.

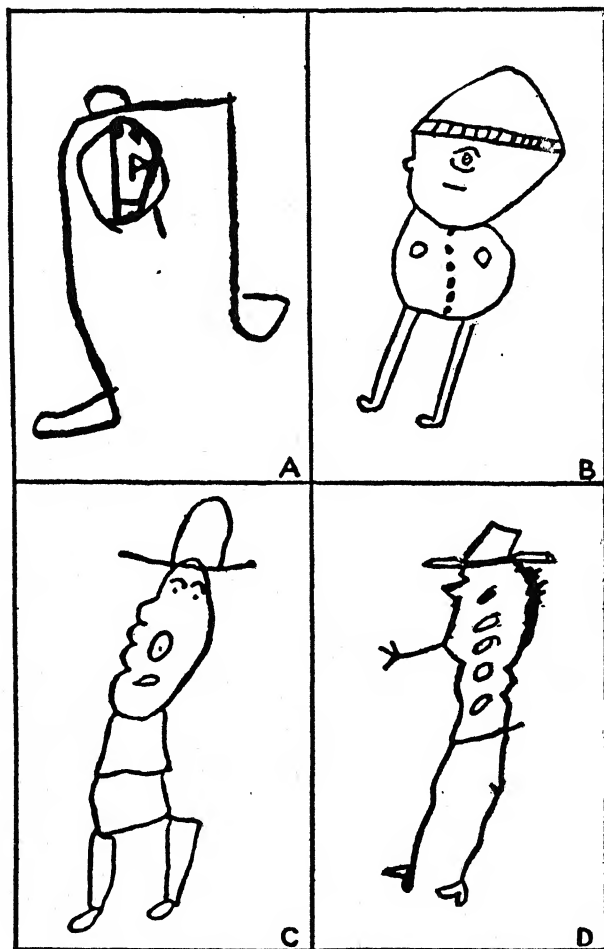


FIG. 42. DRAWINGS BY FEEBLEMINDED CHILDREN. (From *Measurement of Intelligence by Drawings* by Florence L. Goodenough, courtesy World Book Company.)

By means of the development of instruments based upon signs rather than upon samples, many if not most of which might be noted in the course of a routine psychometric examination, not only may it become possible to effect significant improvement in diagnosis and prediction,

but important additions to our scientific understanding of the phenomena that we are attempting to measure may also be gained.

DELINQUENTS AND CRIMINALS

Little need be added here to what has already been said in Chapter 32. We may note, however, that in spite of the many comparative studies of delinquents and nondelinquents, criminals and noncriminals, that have been carried out in the past, none seem to have got at the real heart of the problem. A suggestion made more than twenty years ago by Burt (1925) is worth further investigation by more efficient statistical methods than were then known. In a comparison of delinquent and nondelinquent boys, in which many adverse factors, some of which pertained to the personal-social traits of the boy himself and others to his surroundings, were studied, Burt found that while there was no single factor that did not sometimes appear in both the delinquent and the nondelinquent groups, the total number of such factors was far greater among the former than among the latter. For the delinquents the total number noted was between nine and ten per child; for the nondelinquents the average was about three per child. Burt's study is the more significant because his cases were so carefully matched with regard to their general surroundings. The two hundred delinquents and the four hundred nondelinquents who comprised his comparison groups were of the same sex (all boys), the same ages, and similar social class. They lived in the same neighborhood (usually on the same street) and attended the same schools.

Burt's theory that there are quantitative limits, determined by innate factors, to the amount of external stress that any individual can successfully withstand is worth considering. If expressed in the usual type of T-score units with the mean set at 50 and the standard deviation at 10, then an individual whose power of resistance is 70 would be able to stand up without serious mishap under environmental conditions that are likely to contribute to delinquency even though these conditions are rather marked, reaching, let us say, to a scale value of 67-69. But another, whose resistance is less, say only as much as 40, is likely to become delinquent even under average conditions.

What is needed in the study of delinquent groups is something beyond the mere itemizing of single factors. We need to develop scales in which these factors are combined. A good deal is already known about the characteristics of the individual and the features of the environment which make for delinquency, but either most of the investigations carried on so far have been so dominated by the idea of a single contrib-

uting factor that the investigator has been blind to all others, or else they have been conducted by the "shot-gun" method in which all factors that chanced to occur to the investigator have been studied, but with no reference to any principle of organization that would make it possible to see the parts in relationship to the whole.

SEX DIFFERENCES

The scientific study of sex differences has two rather distinct aspects. First, there is the relatively simple question of determining the type and extent of the psychological differences between man and woman. Of these, which are many and varied, an unusually good account has been given by Scheinfeld (1943) that makes it unnecessary to go into this phase of the subject here. Readers who are interested should consult Scheinfeld's book.

Second, there is the question of the extent to which a given person resembles or diverges from the physical, mental, and personality characteristics that are typical of the sex to which he or she belongs. Such characterizations as the "masculine type of woman" or the "effeminate" or "sissy" man bear witness to the general recognition of the fact that not all persons of a given sex are equally like the typical member of that sex in their personal traits.

At least four scales for measuring masculinity-femininity attributes are available at the present time. The Strong Test of Occupational Interests includes a key for this purpose, as does also the Minnesota Multiphasic Inventory. Terman and Miles (1936) have published a separate test of masculinity-femininity, and Goodenough¹⁴ has devised a similar key based upon free association to common words. Each of these measures shows high internal consistency of items within the same scale but only moderate relationship to other scales, suggesting that the particular aspects of the trait that the different devices measure are not wholly the same.

A number of facts of special interest for social and abnormal psychology have been revealed by studies of inversion of the personal-social aspects of masculinity-femininity by means of comparison of those whose scores fall within the range of those ordinarily covered by the opposite sex with those who earn scores resembling those of the sex to which they belong. Both Goodenough¹⁵ and Terman and Bottenwieser (1935) found that divorced women average considerably more "masculine" on these scales than do women in general. Terman and Miles also found

¹⁴ This key has not yet been published.

¹⁵ *Science*, 1946; also a more detailed report to be published shortly.

that homosexual men who assume the feminine role in the relationship earn markedly feminine scores on their M-F test. Many other findings derived from these extreme groups have important implications both for the scientific study of cultural phenomena and for clinical diagnosis.

NATIONAL-RACIAL DIFFERENCES

Many investigations have established the fact that the representatives of different European groups who have emigrated to the United States show very unequal facility in dealing with the problems included in the typical intelligence test. The differences found for the various immigrant groups have been very similar from one investigation to another, and are approximately the same for the American-born children of immigrant stock as for their foreign-born parents. In practically all the studies that have been made, the average score earned on standard intelligence tests by the Italians, Mexicans, Portuguese, Poles, and Greeks in the United States have been considerably below those of the Jews, Swedes, Norwegians, Englishmen, and Germans. The Chinese and Japanese groups have also ranked fairly high on these tests. In all such comparisons it is of course assumed that the tests used make little or no demand upon a knowledge of the English language.

The uniformity of the findings by many different investigators leaves little room for doubt as to the nature of the bare facts. The interpretation of these facts is another matter. Certainly, without further investigation it would be highly unsafe to assume that the immigrant groups in question are unbiased samples of the population of the countries from which they came, or that any bias which may have affected their migration was the same for all. As a matter of fact, a number of studies carried out in the countries where these persons formerly resided have failed to reveal any differences in intelligence test performance corresponding to those found among immigrant groups in the United States. It may well be true that the factors making for emigration have varied rather markedly from one European country to another. In some, the trend may have been in the direction of forcing out the incompetents. In others, the conditions may have favored the movement of those with superior vigor and initiative. Thus it is quite possible that the findings of American investigators with respect to the intelligence-test performance of immigrant groups in the United States may have been largely or wholly the result of dissimilarities in the conditions making for emigration in the various European countries, rather than due to actual differences in the average intellectual level of the inhabitants of these countries. It is also possible that the particular tests used

in these studies may be poorly suited to the interests and experience of certain groups, even though they may be well adapted to the majority.

Whether or not the second of these explanations can account for the typically low standing of American Negroes on these tests is a highly controversial issue to which no completely satisfactory answer can be given at present. No one can question the fact that the educational opportunities available to Negroes in the South have been inferior to those provided for White children. Even in the northern states, racial prejudice is likely to debar Negroes from many occupational openings. They must live in segregated areas and conform to certain unwritten but well-defined rules in their intercourse with Whites. Unquestionably such treatment affects the emotional life of the Negro and has a bearing upon the development of his attitudes toward life and his level of aspiration. Some psychologists believe that the facts just mentioned provide a sufficient explanation both for the test results and for the relatively small proportion of Negroes who have made important contributions to scientific knowledge or to social welfare. Others feel that more basic factors are involved.

A detailed account of the many studies of national-racial differences would be out of place here. Tyler (1947) has presented the findings of the more recent and important of these investigations together with a very competent discussion of factors that may account for the results obtained. Those who are interested in the topic will find her treatment of this material well worth the reading.

MENTAL DIFFERENCES AMONG CULTURAL GROUPS

Both psychologists and cultural anthropologists have found that comparative studies of groups set apart by reason of geographic factors, religious beliefs, or unusual social customs can be highly informative with respect to the rules that govern behavior. The great difficulty in these studies, as is true of those having to do with national-racial differences, is the question of cause and effect. For example, persons living in rural areas are in general more conservative in their attitudes than those living in the city, but is this because there is more frequent interchange of opinion between man and man among city dwellers, wider social contacts, more direct contact with a variety of living conditions? Or is it possible that the more restless, ardent, and adventurous among the rural population have tended to emigrate to the city while those who are more stable, contented, and cautious in temperament remain on the farms? That selective emigration of this kind has been an

important factor in bringing about a good many of the differences among cultural groups is well known. The relatively low standing on intelligence tests of the dwellers in certain mountain valleys or "hollows" studied by Sherman and Henry (1933), which they regard as the result of environmental deprivation, can be accounted for equally well on the basis of the selective migration process which they say took place at the time these valleys were settled.

THE PHYSICALLY HANDICAPPED

Pintner, Eisenson, and Stanton (1941) have given a very complete and detailed account of the factual results of a large number of studies on the mental and personal characteristics of physically handicapped children. Their summary deals mainly with the deaf and hard of hearing, the blind and weak-sighted, the crippled, and those defective in speech, with briefer discussions of a number of other groups, such as the diabetic, those suffering from tubercular, cardiac, asthmatic, and other chronic weaknesses, including epilepsy and the aftereffects of encephalitis and meningitis. In reading this report one cannot help being impressed by the absence of originality shown by those who made the original studies, as well as by their apparent lack of insight into the problems with which they dealt. Intelligence as measured by standard tests is the characteristic mainly dealt with, and the symptom rather than its cause forms the basis for the classification. Potentially, much may be learned by comparing the mental and personal characteristics of the physically damaged with those of the physically normal, but only confusion can result when the grouping takes account only of superficial appearances with little or no regard to more fundamental differences. A child suffering from cerebral birth palsy, for example, is no more like one who has lost a leg from an accident than the latter is like one who is physically sound. To classify both under the single head of "crippled children," as many investigators have done, is to obscure the facts. Children whose hearts have been weakened from rheumatic fever, with its frequent accompaniments of chorea and other nervous and systemic disturbances, should not be classed with those suffering from other types of organic or functional ailments of the heart.

A field of inquiry which holds much promise for many aspects of neurology, psychiatry, and psychology is the study of the mental and behavioral effects of organic brain damage. These effects are sometimes grouped together under the general designation of *psychological deficit*. As distinct from *mental deficiency* which, as has previously been pointed out, refers to a condition that has existed from birth or from a very

early age, *psychological deficit* has reference to a *loss* in mental ability, to a regression from a higher to a lower level of intellectual efficiency. In addition to the cases with known organic basis (the deterioration due to old age, general paralysis resulting from syphilitic infection, brain tumors and other lesions of the brain substance including accidental or surgical damage), cases in which no organic defect has been identified but in which there is nevertheless definite loss of mental ability such as occurs in certain types of schizophrenia are also included in this group.

Studies in this field are for the most part confined to the present century. The use of mental tests for estimating the extent of the deficit dates from World War I. Since that time, the possibilities for the investigation of the nature of intelligence inherent in this type of approach have appealed to psychologists, while neurologists and physiologists have been equally impressed by the opportunity thereby afforded for the study of the localization of brain function and other neurological problems.

The great difficulty in the study of neurological deficit arises from the fact that in the majority of cases no objective basis for determining the original mental level of the subjects is available. This, of course, renders the question of *loss* not easy to answer. A number of devices have been proposed for its solution, all of which are based upon a comparison of the subject's performance in areas believed to be less susceptible to deterioration with that in areas for which the deficit has typically been found to be rather marked. Among the methods most commonly used in that proposed by Babcock (1930, revised 1941), which is based upon the theory that the score on a vocabulary test is less affected by brain damage than other intellectual manifestations since it was found to remain comparatively high in patients suffering from practically every type of mental disorder. She therefore devised a series of tests having to do with the immediate acquisition of new material. Scores on these tests were equated with vocabulary scores for normal people in such a way that the average difference between the adjusted scores on the two measures was zero. To this difference she gave the name of an "efficiency index." It has been found that persons suffering from many of the known organic brain defects are likely to obtain efficiency indexes below the normal standard on this test and that a less marked difference in the same direction is usually found for the schizophrenic group as well. That this is true of the average patient has been demonstrated not only by Babcock herself but by others who have used her method. Nevertheless, as a result either of differences in the pattern of intellectual ability or of the experimental error of the test itself or of both, a good deal of overlapping occurs between the

normal and the abnormal populations. While the Babcock test may be looked upon as a useful aid to diagnosis and classification, it cannot be regarded as an infallible sign that deterioration has occurred.

Many other theories concerning the nature of the deficit resulting from various types of cerebral deterioration or injury have been proposed, and a number of ingenious tests have been worked out on the basis of these theories. Of these, Goldstein's (1939) contention that organic brain damage is always accompanied by a loss in what he terms "abstract" behavior has attracted much attention. In collaboration with Scheerer (1941), Goldstein has presented a series of tests designed to identify persons in which such loss has occurred. These tests involve the reproduction of designs with colored cubes and sticks, sorting tests, and others. Scoring is in terms of the subject's manner of approach to the problem, his apparent perception of what is to be done, his apperception of the object or pattern as a whole. The damaged brain, according to Goldstein, has lost some of its original power to organize and integrate experience on the basis of abstract concepts of relationship.

From the foregoing brief summary, some idea may be gained of the possibilities afforded by the study of contrasted groups for the testing of scientific hypotheses and for the improvement of our understanding of the laws by which human behavior is governed. Inasmuch as the groups chosen for study usually represent the extremes of characteristics manifested in some degree by most human beings, the results of these investigations have a much wider application than is suggested by a superficial consideration of their overt characteristics. The extreme case throws into sharp relief that which is likely to be overlooked in the more typical instance; it is the magnifying lens by which we are enabled to see what might otherwise be hidden from us. Not merely because the study of the unusual case enables us to deal more wisely with those who are exceptional, but even more because of the better understanding thereby gained of those who comprise the rank and file of everyday human beings, are studies of atypical children and adults important for the scientist.

Looking Ahead

THE ROAD BEHIND US

Much has happened during the four decades that have passed since the publication of Binet's first completely organized scale of mental tests.¹ The little band of psychologists who continued along the path upon which Binet had entered has become a multitude. The path has broadened into a highway from which many trails diverge. Some of these roads were well laid out and are much traveled. Others are still little more than footpaths where the weeds of ignorance grow thickly and the traveler's view is often obscured by the dust arising from his own incautious footsteps.

The road is long and the end is not in sight. Yet as we turn to mark our progress, the beginning of the route seems small and far away. Nevertheless, certain landmarks stand out as reminders of earlier points in the journey. From them we can gauge the distance we have covered; by them we can note the increasing breadth in our point of view.

At the outset we were concerned almost wholly with the measurement of intelligence. Four events of outstanding significance at once attract our attention, since each of these marked the dawn of a new concept and initiated a new method of approach to the problem. First, there were Goddard's translations of Binet's 1908 and 1911 scales into English and his demonstrations of their value for use with American children. Terman's 1916 revision of the Binet, which dominated the field for more than twenty years, introduced the concept of the IQ and presented striking evidence to show that the usefulness of the scale is by no means confined to the identification of backward and feeble-minded children. Equally or even more important is the recognition of those whose ability qualifies them to be the future leaders of their generation. Terman's interest in the child of superior intellectual gifts led to the initiation in 1921 of his justly famous longitudinal study of more

¹ The list of tasks generally known as the "1905 scale" hardly merits that name since there was no attempt at organizing the material beyond a rough arrangement of the items in order of difficulty.

than a thousand cases whose childhood IQ's were exceptionally high, a study that is still in progress.

The next great forward step in the history of intelligence testing was the development of the Army Alpha and the Army Beta tests for use in World War I. This marked the beginning of group testing and thus made it feasible to measure the abilities of the generality instead of the selected few. The introduction of group testing into the schools had an effect upon educational theories and policies that was well-nigh revolutionary. Never before had the extent of the individual differences among school children been so clearly recognized. The folly of treating all alike and expecting the same level of performance from all became apparent, even to the least discerning of school authorities, and many procedures for individualizing the curriculum were tried out.

Most recent of these great landmarks in the history of intelligence testing is Thurstone's work on the analysis of the broad concept of general intelligence into its basic components to which he gives the name of *primary mental abilities*. This is not only a totally different approach to the measurement of intelligence but also a very different concept of the nature of intelligence. Not enough time has elapsed to permit an adequate evaluation of Thurstone's methods as compared to that accorded the more traditional approach initiated by Binet and continued by his many followers both in the United States and abroad, but the radically different point of view and the tremendous amount of careful statistical work upon which Thurstone's techniques are based merit serious consideration, regardless of what the final verdict may be.

The remarkable success of the intelligence test stimulated hopes that other mental functions and forms of behavior might be attacked by the same methods. Tests of educational achievement were being tried out, even before Goddard's translations of the Binet scales appeared, but a new impetus was given to the measurement of school achievement with the advent of the intelligence test. The establishment of clinics for the study of children's behavior problems under the auspices of the National Committee for Mental Hygiene during the early 1920's revealed as never before how such problems were created by children's difficulties in the subjects of the school curriculum. Diagnostic tests intended to show not merely the existence of educational deficiencies but also their specific nature were developed and used as a basis for remedial teaching. Tests for the elementary school were soon followed by tests for high school and college subjects.

A further development in this area is found in the tests of achievement and the tests of aptitude for learning the kind of material covered in the different subject fields, particularly those of the high school and

college where differentiated curricula call for educational guidance if wise choices are to be made. At the earlier ages the so-called "readiness" tests, particularly tests of reading readiness, have been found useful supplements to teachers' judgments as to the wisdom of admitting doubtful cases to the first grade.

Allied to the educational tests are the many and varied tests of proficiency in the vocations of adult life. Industrial concerns have found these tests extremely valuable as guides to the selection and classification of workers. Although many tests of aptitude for the different vocational fields have been devised, it is questionable whether those intended to be independent of the results of specialized training and experience are as valuable as those based upon the theory that aptitude for a type of work makes for greater interest in it, with the consequent probability that some experience will have been sought and gained.

Dating from the moderately successful use of Woodworth's Personal Data Sheet during World War I, attempts to study personality difficulties and emotional disturbances by means of the replies given to a series of selected questions have become increasingly popular in spite of conflicting evidence as to the validity of such reports. Other approaches to the same question as well as to the more general problem of assaying the personal-social traits of the average person have been tried. Among these the various devices for measuring interests and attitudes have proved especially valuable, both as methods of gaining information about the traits directly concerned, and also as indexes of other more general personality characteristics.

The most recent and popular concept of the way to personality measurement is that of projection. Each individual, so it is assumed, "projects" his own inner feelings, his attitudes, desires, and anxieties, his angers, and his fears upon the external world. They are revealed to those who can read the signs in his perceptions of the things by which he is surrounded and consequently in his verbal and behavioral responses to those perceptions. Such tests as the Rorschach and the Thematic Apperception Test are among the most popular of those based upon this concept.

These are but a few of the ramifications that have grown from the apparently simple but extraordinarily pregnant beginning made by Binet during the first decade of the present century. From an instrument designed chiefly for the purpose of identifying the mentally backward there has been developed a series of methods and devices for measuring with greater or less accuracy almost every type of mental process and well-nigh every form of overt human activity that is known to science. From a device with no pretensions to other than practical

applications, instruments of sufficient precision to warrant their use in many types of scientific investigation have been developed and their value for such purposes has become thoroughly established.

Paralleling the development of new instruments and making this development possible has been a corresponding growth in the technical understanding of methods of test construction. Had it not been for the statistical information and methodology provided by such men as Pearson, Thorndike, Yule, Kelley, Thurstone, and R. A. Fisher, the mental test as we know it today could never have been developed.

SOME UNANSWERED QUESTIONS

Widened horizons bring many hitherto unseen points into view. Some, at first but dimly glimpsed, stimulate interest but do not awaken controversy as to their nature or as to the manner in which they may be approached. But as their outlines become more clearly delineated, such questions are certain to arise.

With the rapid multiplication of testing methods and the recognition of the wide range of problems that may be attacked by their use, conflicting points of view were doubtless inevitable. Many factors have contributed to the intensity of these conflicts. Investigators have rushed into print with results obtained for a small sample of cases and have thereafter felt that their own prestige demanded the substantiation of conclusions too hastily formed. Sometimes faulty statistical procedures have befogged the issue. With the refinement of statistical methods, the necessity of establishing basic controls becomes more rigid, and as these methods become more elaborate and complex, the difficulty of gaining a thorough understanding of their underlying assumptions and requirements increases in geometric ratio.

Certain types of statistical error occur and recur. Calculations of probability in which the methods appropriate for large samples were used for very small numbers of cases were understandable enough before sampling distributions had been completely worked out, but even today we find that the likelihood of chance variation is frequently underestimated from neglect of the rules of small samples. Although the effect of errors of measurement upon the correlation of initial scores with gains was pointed out by Thorndike and a corrective formula was derived by Godfrey Thomson (1924) a quarter of a century ago, many people appear to find the principle not easy to grasp, especially when the method used involves some small variation from the conventional method of reckoning correlations. Because a number of the basic questions in the field of mental measurement have to do with the analysis

of causative factors and the associated problems dealing with the development of methods whereby the intellectual or personal-social characteristics of an individual may be improved, the measurement of gain under various types of experimentally interposed factors becomes highly important. Unfortunately, many of the investigations of this topic have neglected to take account of the negative correlation between gain and initial status, with the result that the claims made for the effect of various types of educational training in raising the IQ have vanished into thin air when the proper statistical methods were employed.

The question of the extent to which intelligence can be improved by training is one of the most vexing that has appeared in the entire field of mental testing. Dozens of studies have been made, utilizing various methods of approach. Each of these methods has its own possibilities of error and its own limitations in demonstrating the facts and drawing conclusions from them. Perhaps it is not surprising, therefore, to find that although a quarter of a century has elapsed since the first of the major investigations was begun, the issue is still joined. No final answer can be given to the matter at present, but a few general statements can be made that may help to clarify our thinking.

1. The question as to the possibility of improving the IQ cannot be answered in the negative. Such an answer is forbidden by the laws of science, for we should never forget that a result impossible to present ignorance may become not only possible but even commonplace as ignorance gives place to knowledge.

2. The generalized answer in the affirmative is likewise neither scientific nor practical. We need specific facts. For similar reasons, small attention is merited by those who claim to have accomplished positive results but who are unable to describe their methods with sufficient clarity to make it possible for others to follow them. Particularly when apparently similar methods in the hands of other workers fail to bring about a dependable change in the standing of the subjects is it important to ascertain whether the results actually obtained by the unknown method or the manner of handling the data has been the primary factor responsible for the difference.

3. A test is either a sample of the elements comprising a given universe or a sign from which the character of the universe may be inferred. A sample may be biased; a sign may be misleading. If the universe comprised under the designation *general intelligence* is represented by all the operations demanding abstract thought, judgment and reasoning, memory, imagination, and other mental processes of which mankind is capable, and a *mental test* is looked upon as a series of tasks selected at random from that part of the general universe corresponding

to the type and level of abilities possessed by the class of persons for whom the test is designed, then the validity of the *test performance* of a particular person, considered as an indicator of his level of general intelligence, hinges upon the justification for regarding him as a legitimate member of the group just mentioned. Remembering that the tasks included in the test constitute an exceedingly small sample of the universe from which they were chosen, and remembering further that little interest attaches to the sample if it does not represent the universe, it follows that a subject whose experience has been markedly different from that of the generality of the group with whom he is classified may in some cases no longer be looked upon as a proper member of that group. The standards by which his mates are judged may no longer apply to him. This is a possibility which must not be overlooked when considering the effect of special training upon the intelligence quotient. The training may have affected the sample, i.e., it may have resulted in a genuine improvement in the child's test performance while the effect upon the remainder of the universe of which the test is a sample has been negligible.² Other evidence, not directly associated with the test, must be provided if circular reasoning is to be avoided.

4. As in all such questions where the null hypothesis is involved, the burden of proof rests upon those who take the affirmative side. It is not possible to prove the negative argument, but a positive claim, while its soundness may and should be challenged, is subject to experimental verification. Such verification, however, must be specific. The hypothesis to be tested must be clearly stated in objective terms and the conditions under which it may be expected to hold good must be described in sufficiently precise detail to make it possible for others to reproduce them. Finally, when, as in the case we are now considering, the question has to do with a general ability or form of behavior of which the performance on a particular test is assumed to be a sample, there must be clear understanding of the distinction between the sample and the total. It is one thing to change the sample. It is another, and usually much more difficult, thing to bring about a corresponding change in the total of which the sample is a part. Much of the confusion that exists at present with respect to the modifiability of intelligence has arisen from just such a semantic error.

The controversy initiated by Spearman in 1904 on the question of intelligence versus intelligences is still unsettled, though the heat of battle has diminished. Nevertheless it is a far cry from McNemar's finding that a single general factor can account for nearly if not quite all the

² Since the sample is a part of the universe, at least an infinitesimal effect of a change in the former upon the total into which it enters must be assumed.

variance in performance on the 1937 Stanford-Binet, to Guilford's recent discovery of twenty-seven factors needed to account for variability of success in the army aircraft training program, only two of which were not regarded as "abilities." It is true that aptitude for aviation is not entirely the same thing as general intelligence in the sense that Terman has used the term. Nevertheless, one would hardly expect so great a difference in the number of underlying abilities as those just reported. Thurstone's factorial analysis of his measures of "general intelligence" into twelve primary mental abilities is another case in point.

The basis for these differences appears to lie in varying concepts of the nature of intelligence with consequent differences in the kind of test items taken to be samples of it. Until these differences have been resolved, it is futile to inquire whether "intelligence" is a unitary trait or one made up of a number of more elementary factors. The only possible question that can be asked at present has to do with the number and character of the factors that must be postulated in order to account for the variance in the performance of a given group of individuals on some specified test or battery of tests.

Related to this question but not identical with it is the division between the holistic and the analytic view of the personality. To those who espouse the former, neither a "trait" nor a "factor" has meaning except in relation to the total that it helps to create and by which it is itself created. The whole and its parts are inseparable, but the former is primary; the latter are but derivatives of it. The personality sketches drawn up by the skilled Rorschach or TAT examiner are examples of this type of approach.

The opposite view is taken by those interested in factorial analysis and by the makers of specific tests for specific purposes. To them the personality can be adequately described in terms of the elements of which it is made up when each is given its proper weighting in the total complex. The analysts have gone much further in respect to their mathematical theories and procedures than have those who take the holistic position, but the latter have developed their own methods of examining their data and for testing the hypotheses drawn from it.

Mention should also be made here of the topological methods of studying personality and behavior developed by the late Kurt Lewin (1935, 1936). While these methods in their present form are better adapted to the examination of the general rules and principles that apply to all than to the classification of individuals on the basis of their differences, it is nevertheless true that insufficient understanding of the

general case may lead to gross misinterpretations of the particular instance. Although Lewin's work has had relatively little direct influence upon theories or methods of mental testing, it has suggested new points of attack upon old problems and has pointed the way to new and perhaps highly significant views of personality organization.

In spite of these differences in methods and in the theories that gave rise to them, there is far greater agreement with respect to the methods used for studying abilities, skills, and knowledge and for predicting their future growth than is true in the case of those used for studying personal-social characteristics or in the prediction of conduct. A major stumbling block in the way of arriving at sound conclusions as to the relative value of such devices as the personality inventory, the various projective methods, ratings such as are advocated by Cattell, physiological measures and sampling methods based upon actual behavior is the lack of satisfactory criteria by means of which such measures can be evaluated. No greater contribution could be made to the study of personality differences than would be had from a comprehensive examination of the criteria used for such studies and a thoughtful appraisal of their adequacy for their designated purpose in terms of their objectivity, stability, and psychological meaning. Even a cursory reading of Ellis's (1946) summary of the reported material having to do with the validity of personality inventories reveals the superficiality of the criteria he used. Nevertheless, more suitable ones are not easy to come by.

Of all the unanswered questions in the field of testing, of which those mentioned here represent only an exceedingly small fraction, none is more important and pressing than that of the measurement and appraisal of the personal-social characteristics of an individual. Because we are not dealing with the question of abilities, where the assumption that desirability runs parallel with amount is at least reasonable, a distinction must be made here between the quantitative measurement of a trait and its evaluation in terms of some social, ethical, or industrial standard. The complexity of the problem is readily seen if one considers a single example. At what point on the continuum running from self-confidence to dependence does the optimum fall? Certainly not at either extreme. Nor is it likely that it will be the same for all. It may vary somewhat with sex; it will certainly change with age. Different occupations call for varying amounts of it. Too much would be a handicap for some; too little would spell failure in others. Evidently the answer cannot be a simple one, but its importance is such as to merit careful study by the best minds the profession can offer.

WHAT OF THE FUTURE?

That mental testing has amply demonstrated its value is apparent in the changed attitude of the public. Previous to World War I, few outside the field of psychology had even heard of it. Even among psychologists, skepticism was encountered on every hand. In more progressive schools where some testing was done, parents not infrequently became very indignant if they learned that their child had been given a test. To them, this was tantamount to an insult, for it implied a suspicion that something was wrong with the child's mind.

That time is now safely behind us. The public may have many wrong conceptions about the nature of mental testing but its interest is keen. Few city school systems are now found in which some testing, at least, is not done. Most large cities employ one or more school psychologists. Some have their own school clinics where children with special handicaps or those showing poor emotional or social adjustment are sent for special study. Hospitals, social welfare agencies, and courts of law, industrial plants and commercial organizations, private physicians, clergymen, and lawyers make frequent requests for mental testing of their clientele. Many of these employ their own psychologists.

The faith which our predecessors have inspired creates a tremendous responsibility for those who carry on their work. We cannot afford to take advantage of this trust by claiming knowledge that does not exist, by assuming an air of authority as a cloak for ignorance, and by making dogmatic pronouncements when only probabilities should be considered.

That mental testing has been vastly improved since the days of Binet there can be no doubt; that infinitely more remains to be accomplished is likewise beyond question. Even the best of our present instruments involves a considerable possibility of error; even the simplest is by no means foolproof. Every device employed for mental measurement has its own set of implicit assumptions, its own limits of applicability, its own hazards of interpretation. Neglect of these factors through carelessness or ignorance on the part of the psychologist who makes use of tests for the guidance of human lives may have consequences nearly as serious as those resulting from similar neglect on the part of the physician who deals with their bodily ailments.

What of the future? Will the coming two score years witness an advance in knowledge and an improvement in techniques of research and measurement at all comparable with that which took place between 1908 and 1948?

I am hopeful that it may. Undoubtedly the changes will be less spectacular, for the difference between failure and partial success always

appears to be greater than that between a mediocre and a good performance. But the present emphasis upon providing better facilities for the training of clinical psychologists, together with the accompanying movement toward demanding certification for those who are to enter the field, is certainly a step in the right direction. For the future of mental testing, like that of any other art, depends upon the scientific training and insight of those who are now entering the field. To them we must look for new knowledge and wider views. The future is theirs to create. I am confident that the work will be good.

Glossary

(Italicized words or phrases used in the definitions are defined elsewhere in the Glossary.)

A

abnormal As distinguished from "subnormal," the term refers to a qualitative departure from the normal, a difference in kind, rather than in degree or level of maturity.

abscissa The base line of a chart or graph.

absolute scaling The process of developing a system of measurement in which the distances from one unit to the next are equally spaced with respect to some specified attribute (such as difficulty) and in which the position of the zero point is known.

accomplishment quotient (AQ) also known as the **accomplishment ratio (AR)** A method proposed by Franzen for making a quantitative comparison between the educational standing of a child and his level of intelligence. The procedure, which has since been shown to be invalid, consists of dividing a child's educational age by his mental age.

achievement test A test designed to measure the extent of knowledge or skill that has already been attained in a given field.

action current An electrical current which accompanies a wave of excitation in a nerve, muscle, or gland and which is observable on a galvanometer as a negative deflection. (Warren, 1934.)

adjustment questionnaire A set of questions designed to uncover the personal and emotional difficulties which an individual is experiencing. Also known as a personality inventory.

affective Having to do with feeling or emotional tone.

age, chronological Age reckoned from the date of birth.

conceptional Age reckoned from the date of conception.

developmental See *developmental age*.

educational See *educational age*.

mental See *mental age*.

age scale A scale in which the items are arranged in groups according to the age at which a certain proportion (usually about 70 per cent) of children are able to pass them. Also known as a year scale.

altitude of intelligence A term used by Thorndike to indicate the average difficulty of the tasks which an individual is able to perform.

ambidexterity The absence of definite hand preference; ability to use one hand as well as the other.

ament A mentally defective person.

analogies test A test in which the subject is required to choose from a list of terms that one which is related to the stimulus word in a manner corresponding to that of a given model as in the following example:

Grass: green. Sky: high, clouds, blue, heaven.

Numbers or pictured objects are sometimes used in place of words.

analysis The division of a complex universe into its simpler components.

of variance A method of determining the likelihood that each of a series of samples was drawn from the same universe. It is based on a comparison of the estimates of the variance of the universe made from the variance of the means of the samples with that based upon the variance of the cases included within the samples.

anthropology The science which deals with the study of mankind. Broadly speaking, it includes all the sciences having to do with the structure and behavior of the human organism, but it is mainly concerned with group tendencies rather than with individual differences.

anthropometry The science which has to do with the physical measurements of the human body.

appraisal The application of a system or scale of values to an object, an individual, or an action. Appraisal differs from *estimating* or *measuring* in that it implies the attachment of some standard of worth or importance to the results obtained.

aptitude test A test designed to predict success in a given field in advance of actual trial.

area of intelligence A term used by Thorndike to designate the total intelligence. It is shown graphically by plotting the successive *ranges of intelligence* at each level of *altitude*.

array The distribution of scores on one of two correlated variables corresponding to a given score on the other.

artifact An artificial consequence of the manner in which data are secured and handled rather than a principle inherent in nature.

assaying (1) A term not widely used in mental testing, which refers to a test or experiment designed primarily for the purpose of giving appropriate weight to certain parts or aspects of a complex; (2) loosely used as a synonym for *appraisal*.

aspiration, level of A term used to indicate the degree of success that an individual hopes and plans to gain; the standard that he sets for himself along any specific line of achievement.

associationism A psychological theory or system that traced all mental operations, no matter how complex they might be, to association

with previously experienced situations, going back eventually to simple sensations.

attenuation In its statistical sense the term refers to the reduction in the magnitude of the correlation between two variables resulting from the experimental errors of the measurements.

attitude A stabilized set or disposition.

attitude scale A scale designed to show in quantitative terms the degree of favorableness or unfavorableness of regard for a specified person, group, or social institution.

audiometer An instrument for measuring auditory acuity at different pitches.

average A measure of the central tendency of a group of scores or measurements. In popular language, the term "average" is frequently identified with the *arithmetic mean*, but in a more general sense it is applied to any single statistic used to represent the group as a whole. The *arithmetic mean*, the *median*, the *mode*, and the *harmonic mean* are the averages most often used.

average deviation (A.D.) A measure of variability found by subtracting each of the individual measures in a *frequency distribution* from the mean of the distribution and finding the mean of these differences without regard to their sign.

B

basal age The highest year level at which a child is able to pass all the items of a Binet test.

battery, test A series of tests standardized on the same group of subjects and planned to be administered as a unit but covering a variety of subjects. A battery of educational tests, for example, would include tests of proficiency in all the major subjects of the school curriculum, each administered and scored separately, but all given on the same occasion. Usually a method is provided for combining all the tests of a battery into a total to which an appropriate generalized name is applied.

bias Distortion which tends to take a particular direction. An opinion is biased if it leans always or usually toward a particular system of belief or attitude; a *sample* is biased if it includes an undue proportion of certain elements in the *universe* for which it stands with consequent underrepresentation of others.

biometrics (biometry) The science which deals with measurements of living organisms.

biserial r A method of computing *correlation* when one of the *vari-*

ables is *continuous* and the other, though assumed to be truly continuous and *normally distributed*, is constricted into two broad classes.

blind analysis A method of testing the validity of tests by matching descriptions of the subjects. One set of descriptions is prepared by the examiner and is based wholly on the test results. The other is prepared by a second person and is based on acquaintance with the subjects. The test of validity depends on how well a third person can match the individual descriptions in the two sets.

Brace test A test of motor abilities for adolescents, named for its author.

Broca, region of The part of the brain chiefly involved in the control of the mechanisms of speech.

C

calibration The process of rectifying the units on a graduated scale in accordance with some uniform system of values. In test construction, the term has particular reference to the process of transforming the results obtained by a mere counting of items passed to equally spaced units of measurement.

category A class of objects or conditions in which no attempt at further quantification has been made.

cause A factor that precedes another in time from which the occurrence or nonoccurrence of the second can be predicted with a degree of probability that is significantly greater than chance.

ceiling (of a test) The highest score that can be made on that test; by implication, the highest level of ability that the test in question can indicate.

Chi-square (χ^2) A *statistic* used to determine the probability that a given system of deviations within a series of *samples* can be reasonably supposed to have arisen by chance if all were taken from the same *universe*.

cluster (statistical) A group of *traits* psychologically very closely related to each other and well differentiated from the remainder of the series studied. The grouping as determined by a *factorial analysis*.

coefficient, of intelligence A method of expressing the result of an intelligence test derived by dividing the number of score points earned by an individual by the number that is the standard for his chronological age. Except that, as originally proposed, the data were not equally *calibrated*, the CI is identical with the *Heinis PC*.
of mean square contingency See *contingency coefficient*.

concept A generalized idea based upon knowledge of the qualities common to more than one object of a class; an abstraction.

conceptional age See *age, conceptional*.

congenital Existing from birth.

consistency, internal See *internal consistency*.

constancy of the IQ The theory that the *intelligence quotient* of an individual child remains the same (within the limits of errors of measurement) from early childhood until the onset of senility.

contingency coefficient A method of determining the relation between two series of data chiefly employed when the findings are expressed in *categorical* rather than *continuous* numerical terms, e.g., the relation between state of residence and average amount of schooling.

continuous A term characterizing a series theoretically capable of infinite subdivision (such as length or time). Contrast with *discrete*.

correlation The tendency for two measures to vary concomitantly.

Spearman's rank-difference method A procedure for finding the *correlation* between two series of *variables* by a comparison of the differences in their rank order. Used mainly when the number of cases is small.

correlation matrix A tabular arrangement of all the intercorrelations of a series of *variables*, in which the same order of placement is followed for both rows and columns.

multiple The *correlation* between the combined scores on a series of measures and a *criterion* when each of the measures in the *battery* is given a weight which will yield the highest possible correlation with the criterion for the battery as a whole.

partial The degree of relationship between two *variables* that remains after the effect of one or more other variables, common to both, has been removed by reducing the *variance* of these factors to zero, thus assigning them a constant value for all cases.

correlation ratio A method of determining the relationship between two *variables* when the lines of *regression* are not *rectilinear*.

correlation surface See *scatter diagram*.

counting An enumeration of *discrete* objects or events.

cretin A mentally defective person whose condition is due to inadequate secretion of the *hormone* from the *thyroid gland*.

criterion A standard used in checking the meaning or nature of a test or sign. The *validity* of a test, that is, the extent to which it measures whatever it purports to measure is judged by the extent to which it agrees with an accepted criterion.

external A standard of reference determined without direct reference to the measure with which it is to be compared. When one part of a test is used as a standard for another part, this is sometimes called an "internal criterion," but the term *internal consistency* is more appropriate.

critical score A level of test performance below which experience has

shown that success in a particular situation is so unlikely that it is unwise to advise candidates to undertake it.

cross-sectional method In contrast with the *longitudinal* method of studying developmental processes, the term refers to the use of presumably representative *samples* drawn from successive age groups of the population. In a more general sense the term is used with reference to any *random* drawing from a given *universe*.

culture-epoch theory A special aspect of the *theory of recapitulation* which held that the progress of primitive man toward civilization is mirrored in the developmental changes in the play of children.

curve In statistics, (1) the line formed by connecting the points determined by the heights of the successive *ordinates* in a *continuous* series; (2) the surface determined by such a boundary line; (3) the path of a point which is defined by a mathematical equation as the *curve of probability*.

cycloid Oscillating between depression and elation.

cyclothymia A condition in which the subject shows excessive changes of mood either without observable cause or in response to relatively trifling circumstances; in extreme cases leading to *manic-depressive* attacks.

D

D The 10-90 per cent *range* in the distribution of a *continuous variable*.

decile The range of scores covered by any single division of a serially ordered group which has been divided into ten equal parts.

decile points The points marking off one decile group from another.

deficiency, mental (1) A state of mental backwardness or retardation as compared to the generality of persons of similar age; (2) commonly used as a synonym for feeble-mindedness, indicating a degree of intellectual retardation existing from birth or from an early age so profound as to render the individual unable to compete on equal terms with his normal fellows or to manage himself and his own affairs with ordinary prudence (Tredgold).

definition, operational See *operational definition*.

delinquency area A part of a city in which significantly more juvenile delinquents reside than is to be expected on the basis of its population. The term may also be applied to rural areas in which the same condition exists.

delusion An error of judgment or reasoning, a mistaken belief. (See *illusion*.)

systematized A series of delusions organized about a common object or situation.

- dementia praecox** See *schizophrenia*.
- dependent variable** See *variable, dependent*.
- desurgency** The opposite of *urgency*; a tendency to look on the dark side, to be worried or depressed by small matters.
- developmental age** A broader term than *mental age* since it refers to the level of development of the entire *personality* rather than to *intelligence* alone. Specifically it is used in connection with the results of tests given to infants and with those obtained by a test developed by P. H. Furfey which deals particularly with the maturity of the interests and attitudes claimed by the subject.
- deviation, average** See *average deviation*.
- standard** See *standard deviation*.
- dichotomy** (1) The process of dividing an object or a series into two parts, not necessarily of equal size: (2) one of the parts so formed.
- point of** The point on the base line of a distribution curve at which the division is made.
- differences, individual** See *individual differences*.
- discontinuous** As applied to a series, the term refers to items which are qualitatively distinct from each other so that the limits of the classes cannot be set at will but must conform to the nature of the items. A discontinuous (*discrete*) series can be handled by counting, but the series as a whole cannot be *measured*. Contrast with *continuous*.
- discrete** Constituting a separate object or category, not belonging to a continuous series. Contrast with *continuous*.
- discriminative value** Literally, the worth of a measuring instrument for making distinctions between individuals or groups. Technically, as used by Arthur and Woodrow, the D.V. is the difference between the *mean* scores made by two consecutive age groups divided by the *mean* of the *standard deviations* of these groups.
- disease, mental** Any serious disturbance of mental functioning of a more or less lasting character which is first manifested after a period of normal or at least more nearly normal mental status. Now more often known as a psychosis (pl., psychoses), usually differentiated as major or minor according to the extent to which the stated condition is likely to disrupt the patient's normal life as a member of society.
- dominance, lateral** See *lateral dominance*.
- dynamograph** A recording *dynamometer*.
- dynamometer** An instrument used to measure muscular force. The most common form is the hand dynamometer (see Figure 29), which registers the strength of grip.

E

- eclecticism** The selection of concepts and principles from a variety of schools of thought accordingly as they seem to yield consistent explanations of observed phenomena.
- educational age** A semiquantitative score obtained in the same manner as is the mental age except that tests of school achievement are used instead of intelligence tests.
- educational quotient (EQ)** The ratio between a child's *educational age* and his *chronological age*. (The term should not be confused with the *accomplishment quotient*.)
- entelechy** The result of a formative or organizational process whereby a higher or more complex entity is established.
- episode sampling** A method of measuring certain forms of behavior in terms of the number of times in which it is observed during a specified period of time
- equal-appearing intervals, method of** A method used chiefly in the development of *attitude* scales. It is based upon the pooled judgments of a large number of persons to the effect that the intervals between the items in a scale "appear" equal.
- error, probable** See *probable error*.
- standard** See *standard error*.
- estimate** A judgment, usually of a *quantitative* or semi-quantitative nature, which is based upon observation, report, or other data not strictly *quantitative* in nature, or upon a *sample* from which a judgment of some quality of the total is to be made.
- eta (η)** The Greek letter used as a symbol for the *correlation ratio*.
- ethical judgment test** A test in which the subject is called upon to make judgments concerning the moral or ethical qualities of certain described acts.
- eugenics** The scientific study of methods for improving the racial stock.
- experiment** As distinguished from a *test*, a *measurement*, or an *investigation*, an experiment involves the setting up of a specific question or series of questions of general scientific import and the arrangement of a series of varying situations such that logical inferences can be drawn from the nature of the responses of subjects to changes in the experimental conditions.
- extrapolation** The process of extending the range of a series of standards above or below the interval over which measurements were actually taken, by means of mathematical calculation of the trend of the curve.
- extraversion** Tendency to direct the attention toward external things and events without marked awareness of the self; relative absence

of introspection, concern about imagined happenings, or dwelling upon personal experiences. Contrast with *introversion*.

extrinsic Pertaining to factors outside the person or thing in question. Opposed to *intrinsic*.

eye dominance The tendency shown by most persons to "sight" with one eye rather than with the other.

F

F A *statistic* used in estimating the probability of exceeding a given divergence in the *variance* of two *samples* which takes account of the difference between the variance of a *sample* and that of its *universe*.

factor (1) One of the elements or quantities which enter into a product as determined by *factor analysis*; (2) a condition which, in combination with others, operates to bring about a given result.

factor (factorial) analysis The statistical analysis of the intercorrelations between the results of a number of tests or measurements presumably representing a given mental function (such as the separate subtests of a group test of intelligence), with the purpose of ascertaining how many separate part-functions or *factors* it is necessary to postulate in order to account for these relationships and thus, presumably, to describe in the simplest possible terms the organization of the *trait* as a whole. A number of different methods of factor analysis have been proposed, according to the various theories of mental organization held by their authors.

faculty (1) A term popularly used to denote almost any mental ability; (2) historically, in terms of *faculty psychology*, it was used to denote a specialized power or agency of mind through the action of which certain types of behavior were made possible.

faculty psychology A system of psychology, popular in the past, which was based upon the classification of mental processes under a relatively small number of heads which were treated as distinct entities with a more or less fixed system of organization, and on the basis of which explanatory principles for observed facts of behavior were drawn up.

fallacy, naming See *naming fallacy*.

fiducial limits The limits set by any specified level of probability.

figure As used by the school of *Gestalt psychologists*, the term applies to a unified impression, derived from a single sense, which, to the observer, appears to "stand out" as something distinct from the relatively undifferentiated background or "ground."

floor (of a test) The lowest score that it is possible to make; by impli-

cation, therefore, the lowest level of ability that the test in question can indicate. Contrast with *ceiling*.

forecasting ability, index of A statistical formula from which the average per cent of improvement over sheer guess that is made possible by using the scores earned on one test or measurement as a means of predicting the most probable score on a second is determined.

freedom, degrees of The number of free choices that can be made under a given set of experimentally imposed conditions.

frequency distribution A table, chart, or graph arranged to show the number of cases at each successive point in a series.

function (math.) A quantity of which the value in any particular case is determined by the value of one or more other *variables* to which it is related.

G

g The "general intelligence" factor according to the theory advanced by Spearman.

galvanometer An instrument for measuring the strength of an electric current.

gene A factor located in the germ cells that is assumed to be involved in the transmission of a particular *trait* from parent to offspring.

genetic Originating in the genes, that is, hereditary and constitutional in nature.

genetics The scientific study of heredity.

Gestalt, psychology of The school of psychology which holds that all experience occurs as forms or structures (German "Gestalt [en]") which are functionally indivisible and if incomplete tend immanently toward their own completion.

H

hallucination A disorder of perception in which objects or situations not actually present to the senses are experienced and reacted to as if they were present. Contrast with *illusion*.

halo effect The tendency for judgments of one *trait* to be influenced by a knowledge of the subject's standing on other traits or by a general impression with respect to his superiority or inferiority.

"haptic" type of graphic expression As used to Löwenfeld, the term refers to an attempt to depict a scene or an object in terms of inner feeling; its emotional connotation for the artist.

Heinis (PC) See *personal constant* (*Heinis*).

heterogeneous Varying in respect to one or more qualities or characteristics.

hierarchy A system of groups or classes arranged in serial order.

- holistic** Having reference to a whole which cannot be divided or analyzed without changing its essential nature although its attributes may be described.
- homogeneous** Similar in respect to one or more qualities or characteristics.
- homograph** A word which, in its written form, has two or more different derivations and therefore two or more distinct meanings.
- homosexuality** Sexual intercourse between persons of the same sex. Persons who engage in this behavior are known as "homosexuals." The adjective form of the word is also used to describe interests or attitudes which suggest homosexuality.
- hormone** One of a group of chemical substances produced by certain bodily organs which, after entering the blood stream, induce functional changes in other organs.
- hypnosis** A condition that may be artificially induced in certain persons during which suggestibility is so greatly increased that a variety of mental and motor acts will be carried out as instructed by the experimenter.
- hypnotism** The scientific study of *hypnosis* and its related phenomena.
- hypochondriasis** Undue worry over matters of personal health.
- hypomania** Excessive emotional excitability, approaching but falling short of actual mania.
- hypothesis** A preliminary assumption usually based upon enough observation to place it beyond the class of mere speculation but which requires further experiment for its verification. Compare *theory*, *principle*, *law*.
- hysteria** An abnormal mental condition characterized by dissociation of certain painful ideas or memories from active consciousness and the substitution therefor of various physical or behavioral symptoms that serve as defenses against the recurrence of the original difficulty. Persons thought to be particularly susceptible to this form of disorder are said to have a "hysterical constitution."

I

- idiot** Originally used to include all degrees of *mental defect*, the term is now generally limited to those of the lowest intellectual levels. In terms of Stanford-Binet test findings, it includes only those persons who at maturity have not attained a mental age exceeding 2 or 2½ years or whose IQ's in childhood do not exceed 20-25.
- idiot savant** A person of subnormal mentality who shows unusual talent or skill in some specialized area such as music, art, or memorizing numbers.

- illusion** A mistaken interpretation of an actual sense impression. Contrast with *hallucination*, in which the abnormal perception has no basis in objective reality, and with *delusion*, which is a broader term referring primarily to a system of ideas about which erroneous judgments are formed.
- imbecile** A feeble-minded person whose level of intelligence is such that he can guard himself against ordinary physical dangers and carry out acts of self-help such as feeding and dressing himself but who can do little or nothing in the way of self-support.
- independent variable** See *variable, independent*.
- index of brightness** One of the early devices proposed for expressing the results of intelligence tests, in which the numerical difference between a subject's point score and the score taken as the standard for his age was to be added to or subtracted from 100 according to its sign.
- index of forecasting ability** See *forecasting ability, index of*.
- individual differences** A term used to subsume the variations in structures or function found when the members of a group of subjects are compared with each other.
- innate** Present at birth. The doctrine of *innate ideas* is the theory, held by some of the older philosophers, that certain fundamental ideas or concepts are common to all persons and hence will arise without previous experience i.e., from "innate" causes.
- insane** A popular and semilegal expression referring to *mental disease*. The term "psychotic" is now preferred to either of the above expressions.
- integral (math.)** The result of integration.
- intelligence** The ability to think in abstract terms (Terman). Also defined as the ability to utilize previous experience in meeting new situations; the ability to make good responses from the standpoint of truth or fact (Thorndike); the ability to reason well or to form sound judgments; the ability to improvise new responses to meet the needs of new situations; etc.
- intelligence quotient (IQ)** The quotient obtained by dividing a child's *mental age* by his *chronological age*.
- constancy of** See *constancy of the IQ*.
- interaction factor** That component of the total *variance* in a series of measures that cannot be accounted for either by the mean *variance* of the individuals making up the groups of which the total is composed or by the *variance* of the means of the groups.
- internal consistency** A term used to indicate the extent to which the separate items or parts of a test are *correlated* with each other.

interpolation The process of calculating intermediate values between measured points in a series.

intrinsic Inherent in the thing itself, not the result of external or accidental factors. Opposed to *extrinsic*.

introversion A tendency toward greater than usual preoccupation with the self; inward rather than outward direction of the attention with consequent introspection often accompanied by some degree of social withdrawal and daydreaming. Contrast with *extroversion*.

inventory, personality See *adjustment questionnaire*.

investigation A term used somewhat loosely to denote the examination of a body of data either collected by the investigator or collated by him from the work of others for the purpose of ascertaining certain facts and relationships. The term "study" is often used as a synonym.

K

K scale A special key devised for use in scoring the *Multiphasic Inventory (Minnesota)* to correct for errors due to malingering.

kurtosis The state of curvature of a bell-shaped *frequency distribution*.

kymograph An instrument for recording small movements by means of a pen or stylus which traces a record on a revolving drum.

L

lateral dominance The tendency to use the eyes or limbs on one side of the body in preference to or more accurately than those of the other side.

law A rule with no exceptions.

leptokurtic A term applied to a distribution of measures in which the massing of scores within the central peak is unusually pronounced, with long tails on either side.

longitudinal Literally, "lengthwise." As used in psychological research the term is customarily applied to studies of the same subjects over a period of time. Contrast with *cross-sectional*.

M

mandible principle The lessened sense of strain incurred when two parts of the body cooperate by opposition as in the action of the jaws in biting, as compared to that felt when an equal force is exerted by the action of a single muscle group working independently.

manic-depressive A term used to refer to a class of mental disorders characterized by excessive shifts from excitement to depression.

manoptoscope (also known as a *manuscope* or *V-scope*): A device for ascertaining *eye dominance*.

manuscope See *manoptoscope*.

mean, arithmetic The result obtained by dividing the sum of the individual measures in a *sample* by the number of cases included in the *sample*.

deviation (M.D.) See *average deviation*.

harmonic The reciprocal of the mean of the reciprocals of the separate measures in a series.

measurement The act or the result of comparing a given quantity in a *continuous* series with a standard scale in order to give numerical expression to its amount or degree. Compare with *estimate*, *appraisal*, *counting*, *assaying*, *analysis*.

error of Differences in the results obtained by successive measurements of a presumably constant attribute of some person or object. Usually expressed as the *standard error of measurement*, which is a statistical *estimate* of the *standard deviation* of an infinite number of such measurements.

median The point which divides into two numerically equal parts the measures of a group arranged in serial order.

median mental age The median of the mental age equivalents obtained for a series of separate tests.

mental age (MA) A semiquantitative term derived by comparing a child's performance on a standard series of tasks with the average performance of children at each succeeding age level until a point is reached to which his own level of success most nearly corresponds. The child is then said to have a mental age equal to the average chronological age of the children at the point of coincidence.

mental defect See *deficiency*, *mental*.

mental disease See *disease*, *mental*.

mental test A standardized task or series of tasks used for the measurement or appraisal of some specialized aspect of ability. Unless otherwise designated, the term is usually understood to have reference to general intelligence.

mesokurtic A term applied to a distribution of scores in which the *kurtosis* corresponds to that of the *normal curve* within the limits of chance variation.

mixed dominance A condition in which the dominant hand is on the opposite side from the dominant eye. See *lateral dominance*. Many clinicians have attempted to trace a relation between this condition and stuttering, but the evidence so far is inconclusive.

mode The point in a frequency distribution at which the greatest number of cases are found.

crude The mode of a *sample*.

- true** The estimated mode of the *universe* from which a *sample* is drawn.
- molar** A term used in psychology to denote the view that the individual person is, in effect, an indivisible unit whose behavior and whose inner life can be adequately understood only as a totality.
- molecular** As used in psychology the term refers to the attempt to construct and understand a whole through an examination of its parts. Contrast with *molar*.
- moron** (from the Greek, meaning "sluggish" or "stupid"): A feeble-minded person whose level of intelligence most nearly approaches normality. Many morons are capable of partial or complete self-support if given adequate supervision, but they are not competent to manage their own affairs or direct their own activities without guidance.
- Multiphasic Inventory (Minnesota)** A questionnaire developed by Hathaway and McKinley at the University of Minnesota that is designed to measure a number of abnormal mental characteristics (hence "multiphasic") on the basis of the symptoms claimed by the subject.
- multiple-choice method** A procedure much used in group testing in which a question or an incomplete statement is followed by a number of answers from which the subject is required to select one on the basis of some specified principle such as correctness, personal preference, incongruity, etc.
- multiple correlation** See *correlation*, *multiple*.
- multiple scoring key** A device for scoring the same test blank according to a number of different *criteria* and thus obtaining data on a number of different characteristics by the administration of a single test.

N

- naming fallacy** Errors arising from the acceptance, usually without conscious thought, of a name given to a measuring instrument as sufficient evidence of the kind of thing which it measures.
- negative acceleration** Progressive decrease in the rate of change in a specified function with the passage of time or the addition of further practice.
- neurasthenia** A condition characterized by persistent feelings of fatigue without known physical cause, accompanied by tendencies to *hypochondriasis* and feelings of inadequacy and insecurity.
- neurosis** A condition in which there is more than the usual tendency

toward nervous tension accompanied by feelings of conflict and unusual difficulty in making decisions.

nondirective interview A method of therapeutic counseling in which the client is encouraged to discuss his own difficulties and experiences freely but the interviewer makes no direct suggestions and gives no advice. The aim is to help the client "think through" his problems and find a solution for himself.

nonsense syllable A group of letters, usually three in number, which can be readily pronounced but do not comprise a meaningful word or word element.

norm A *qualitative* or *quantitative* standard of reference assumed to be typical of a given population and therefore used as a base with which individuals or groups may be compared.

normal Corresponding to the group norm to a degree judged sufficient for most practical purposes. Such deviations as exist are not great enough to warrant classifying the case as either definitely subnormal or definitely abnormal. The term is commonly used in an approximate rather than an exact sense.

normal curve The bell-shaped curve which results when the frequency of occurrence of the successive values in a continuous series arises from the combined operation of a very large number of independent causes. The *integral* of this curve is called the "*normal probability integral*." Tables of this *integral*, showing the numerical values of all the *functions* of the curve at each successive point, will be found in most of the standard textbooks of statistical methods.

normal distribution A distribution that conforms to the *normal curve*.

normalizing The process of converting a series of scores into the form of a *normal curve* by assigning to each the value corresponding to its frequency in a *normal distribution*.

normative group See *standardization group*. • •

null hypothesis A hypothesis framed in negative terms.

O

observation The act of attentive examination of an object or situation.

(1) Observation may be *casual* as when the attention is caught without previous intent and no special effort is made to remember what is seen; (2) *systematic* when made according to some pre-designed plan; or (3) *controlled* when, in addition to such previous planning, special conditions are set up to prevent distractions and to provide an optimal situation for observing certain phenomena.

ontogenetic Having to do with the course of development in the individual. Contrast with *phylogenetic*.

operational definition A definition in terms of an activity or process carried out by an organism or in terms of the results of some specified activity.

“optic” type of graphic expression As used by Löwenfeld, this refers to an attempt to reproduce objects or scenes as they appear to the eye. Contrast with *haptic*.

ordinate The vertical axis of a chart or graph.

overstatement test A test in which the subject is first asked to make a statement concerning his ability to perform some specified task or his knowledge of a particular fact and is later given an objective test to determine the extent to which his claims are in conformity with his abilities.

P

paranoia An abnormal mental state characterized by systematized *delusions*, particularly delusions of persecution.

parsimony, principle of The principle which affirms that when either of two explanations of a phenomenon is possible, the simpler one or the one that demands the fewer underlying assumptions is to be preferred.

partial correlation See *correlation, partial*.

Pearson's r The *product-moment method of correlation* devised by Karl Pearson, which is based upon the sum of the products of the *standard scores* on each of the two *variables* divided by the product of their *standard deviations* multiplied by the number of cases.

“peephole” method A method of studying eye-movements in reading in which the examiner observes the child's eyes through a small hole in the center of a screen on which the material to be read is mounted.

per cent of average See *personal constant (Heinis)*.

per cent placement A method of expressing individual mental test standing in terms of the per cent of the distance between the scores made by the poorest and best of a typical group of 100 subjects of the same chronological age that the individual in question has reached. In contrast with the *percentile rank*, in which the successive units represent equal areas of the distribution curve, the per cent placement is based upon equal distances along the baseline of the curve, and is therefore subject to mathematical treatment.

percentile That point or score in a *frequency distribution* below which a stated percentage of the cases lies; e.g., the twenty-fifth percentile is the point which divides the lowest 25 per cent of cases from the upper 75 per cent.

percentile rank The *percentile* at which the score made by a given

individual falls. Thus, if a particular subject is said to have a percentile rank of 10, it means that 10 per cent of the cases in his group do no better than he on the test in question while 90 per cent surpass him.

perception The direct awareness of some existing external condition that arises through the operation of sensory processes when some degree of integration through experience has given rise to at least a primitive kind of recognition and understanding.

performance test A test requiring the physical manipulation of concrete materials rather than verbal responses.

permille That point on a curve of distribution below which a stated number of thousandths of the cases lies. The term has the same general meaning as *percentile* except that it is based upon a representative group of a thousand rather than a hundred cases.

personal constant (Heinis) A method of expressing test results in terms of the quotient obtained by dividing the score earned on an intelligence test by the score corresponding to the child's chronological age, both expressed in equally spaced units according to a scale derived by Heinis from data collected by Vermeylen. Also known as the "personal coefficient" or as the "per cent of average."

personal equation The name originally given to the differences between two or more observers in their apprehension of observed phenomena.

personality A term used with widely different connotation according to the systematic position of the user. In mental testing, however, one of two definitions is commonly understood: (1) the total integrated character of the individual in his functional dealings with the world about him; and (2) those aspects of his behavior that are particularly concerned with his social and emotional life, that have to do with his conduct as distinct from his abilities.

personality (or personal) inventory See *adjustment questionnaire*.

personality trait A *trait* that has to do primarily with the social or emotional behavior of an individual.

personality type Classification of an individual according to his most outstanding *traits*.

Phi-coefficient A measure of *internal consistency* found by determining the relationship between passing or failing a test item and belonging to the top or the bottom 27 per cent of a group in terms of total score.

phrenology The pseudo science which seeks to determine the mental, social, and emotional characteristics of an individual from the form of his skull.

phylogenetic Having to do with the origin and differentiation of races or species. Contrast with *ontogenetic*.

physiognomy The pseudo science which attempts to diagnose mental and social characteristics from facial characteristics.

platykurtic A term applied to a distribution of scores which is flatter than that of the *normal curve*, i.e., one in which the massing of scores in the center is less pronounced with more of the cases spreading out toward the tails.

plythsmograph An instrument for measuring changes in the size of different parts of the body due to alterations in the distribution of blood.

point scale A scale in which the items are arranged in serial order, usually on the basis of difficulty, and the score is expressed in "points" according to the weighted or unweighted sum of the items passed.

press A term used by Murray in connection with his *Thematic Apperception Test* to denote the kind of external forces which give rise to behavior through inducing a feeling of necessity.

primary mental abilities Abilities regarded as basic in the sense that they involve only a single *factor* and are thus not subject to further statistical analysis.

principle (1) A guiding rule in scientific investigation, as the *principle of parsimony*; (2) a uniform relation in nature on the basis of which other facts or rules may be derived.

probability curve See *normal curve*.

probability integral See *normal curve*.

probable error An *estimate* of the variability of a *universe* based on a *sample* of that universe. Its value is .6745 that of the *standard error*. In a *normal distribution* the range from +1.00 to -1.00 P.E. will include the middle 50 per cent of the cases.

product-moment method of correlation (Pearson's r) The method of correlation generally used when a straight line is the best fit for the *line of regression*.

product scale A series of standards customarily arranged in order of merit for use in assigning quantitative scores indicating the quality of a given type of material product to such things as handwriting, English composition, or some form of manual craftsmanship. The product to be scored is compared with each level of the standard series in turn until one of equal merit is reached. It is then assigned the same scale value as that of the standard.

profile, psychological A graphic representation of an individual's relative standing on a number of mental and physical measurements after these have been reduced to some uniform system of expression.

projective methods A variety of indirect methods for studying the inner life of an individual. The basic theory underlying all these devices is that each person unconsciously "projects" his private feelings and attitudes into his dealings with the everyday situations of the external world and that his actions thus have a symbolic as well as a literal reference. Projective methods are thus aimed at learning to interpret these symbols. Among the most frequently used are doll play, drawing and painting, interpretation of inkblots, puppetry, clay modeling, and other work with plastic materials that permit a wide variety of symbolic structuralization.

prophecy formula (Spearman-Brown) See *Spearman-Brown prophecy formula*.

psychasthenia A condition characterized by obsessive feelings of compulsion and overanxiety about small matters.

psychiatrist A physician who has had special training in the field of mental and behavioral disorders.

psychodrama A *projective method* of studying *personality* based upon psychological interpretation of plays written by the subject or depicted by means of puppets or dolls on a toy stage.

psychogram See *profile, psychological*.

psychologist One versed in the facts, theories, procedures, and practical applications of psychology (Warren). Modern practice tends to restrict the title to those holding the Ph.D. in psychology.

psychometrist A person trained and experienced in mental testing but not necessarily versed in the broader aspects of psychological facts and theories.

psychopathic Pertaining to mental disease.

psychophysics The study of the relations between the physical qualities of an object or situation and the sensations to which they give rise.

psychosis See *disease, mental*.

public opinion poll A method designed to measure the trend of public opinion on various subjects by questioning selected samples of the population, either through direct contact or by mailed questionnaires or straw ballots.

Q

Q The quartile deviation ($\frac{1}{2}$ the range of the middle 50 per cent).

qualitative Varying in kind or quality, rather than in degree.

quantitative Subject to measurement or to counting.

quartile One of the points by which the scores of a group arranged in serial order from lowest to highest is divided into quarters.

Also used to designate the range of scores within one of these specified quarters.

questionnaire An organized list of questions designed for the scientific study of a special problem or series of problems.

R

random According to chance.

range The distance between the highest and the lowest points in a series of measures.

of intelligence A term used by Thorndike to indicate the number or variety of tasks with which an individual is able to succeed at any given level of difficulty.

rank-order correlation See *correlation, Spearman rank difference method*.

reaction time The time elapsing between the administration of a stimulus and the onset of a response to it.

readiness test A test designed to indicate whether or not it is likely to be profitable to begin the training of a child along some specified line. "Reading readiness" tests are among the best-known examples in this class.

recapitulation, theory of The belief that in his development the child repeats the history of the race.

recidivist A person who is convicted of repeated offense against the laws of the state.

rectilinear In or forming a straight line.

reference standard See *standard*.

regression The tendency for the correlates of deviate scores to rank nearer the mean of the *universe* to which they belong than do the corresponding measures in the first *variable*. For example, Galton noted that the sons of exceptionally tall fathers tend to be taller than the average of their generation (correlation) but not as tall, on the average, as their fathers (regression). In like manner, if a group of children who ranked exceptionally high on the first administration of an intelligence test is retested, their average IQ on the second examination will be lower than that first earned, though still, in all probability, higher than the average of the group from which the children were originally chosen.

rectilinear The situation that exists when the *regression equation* is best represented by a straight line. Also known as "linear" regression.

regression equation The equation used for estimating the most probable score on a *dependent variable* on the basis of the score on the *independent variable*.

regression lines Lines of relationship drawn through the means of the successive *arrays* of one *variable* that correspond to each score on the other *variable*.

reification The act or result of reifying.

reify To assign the qualities of objective reality to that which exists only as an abstraction; to think, for example, of intelligence as "something" possessed by man rather than as an abstract noun under which is subsumed certain qualities of his actions. Literally, to reify means to make real that which is not real in the objective sense of the word.

reliability As used in testing, the term refers to the stability of a given measure on repeated applications or, as it is sometimes put, to the extent to which a test is consistent in measuring whatever it does measure.

respondents As used in testing, the term refers to the persons responding to a question or to some other specified stimulus in a formal test or experiment.

rhathymia A condition characterized by a carefree, thoughtless love of adventure and excitement for its own sake.

rho (ρ) The Greek letter used as a symbol for correlations computed by the rank-difference method.

Rorschach Inkblot Test A test in which subjects are asked to state what they "see" in a series of inkblots. Responses are scored according to an elaborate system to which not only Rorschach but many subsequent workers have contributed.

S

s The specific factors required by Spearman's two-factor theory of intelligence.

sample A portion of a total so chosen that the characteristics of the whole may be judged from those of the part with a minimal degree of error.

random A sample selected by methods which ensure that every item in the total has equal chance with every other of being included within it.

reference A group of subjects whose performance is used as a standard with which others are to be compared. Also known as a *standardization group* or *normative group*.

representative A sample chosen, on the basis of previous information about the characteristics of the total, in such a manner that each of these characteristics will be represented in the sample to an extent proportional to their existence within the *universe* as a whole.

- stratified** See *sample, representative*.
- sampling method or technique** The procedure used in choosing a sample.
- scatter diagram** The figure resulting when the scores in one of two related *variables* are plotted along the *abscissa* of a *curve*, with those of the other plotted along the *ordinate*. The number of cases at each level is recorded at the appropriate points of intersection between rows and columns.
- schizophrenia** (also known as *dementia praecox*): A form of mental disorder characterized by dissociated mental processes such as *delusions*, or *hallucinations*, and by inadequate or inappropriate emotional responses.
- schizothymia** A less pronounced manifestation of *schizophrenia*.
- semantics** The science which deals with the history and evolution of word meanings.
- semi-interquartile range** One half the distance between the twenty-fifth and the seventy-fifth *percentiles*. This is also known as the *quartile deviation* (*Q*). The total distance between these points, i.e., twice the semi-interquartile range, includes the middle 50 per cent of the cases.
- sensation** An experience mediated through one of the organs of special sense (the eye, the ear, etc.), which is aroused by stimuli outside the receptive organ and cannot be further analyzed; it is thus the most elemental unit of consciousness.
- sensationalism** That system of psychology, now generally considered obsolete, which looked upon sensations as the basic factors in human experience, from the organization and association of which all more elaborate experiences had their origin.
- sensory** Pertaining to sensations.
- siblings** Children of the same parents. A generic term which refers to brothers or sisters without regard to their sex.
- Σ (upper-case Greek sigma)** (1) The sign of summation; (2) the larger of two standard deviations entering into a statistical formula.
- σ (lower-case Greek sigma)** The sign used to indicate (1) the *standard error* of a given *statistic*; (2) the *standard deviation* of a *sample*.
- sign** (1) That which points toward an objective, a directive agent; (2) a symbol.
- significance, level of** In statistics, the degree of probability that in further samples the results will show a trend in the same direction but not necessarily to the same extent as that found in the single investigation. It is customary to express this probability in terms of chances in one hundred, that is, in percentages. A statement that a

- given result reaches "the 2 per cent level of significance" (or "confidence") means that there are only two chances in one hundred that further investigations under similar conditions would yield results contradictory to those of the first.
- skewness** That attribute of a *frequency distribution* which has to do with its symmetry on either side of the *median*. A *curve* is said to be "skewed" if one of the tails is definitely longer than the other.
- sociogram** A graphic representation of the forms of social organization within a group.
- sociometry** A method of studying social organization in terms of the stated preferences of individuals for each other under various specified conditions. Its chief protagonist is J. L. Moreno.
- source trait** A term used by Raymond Cattell to indicate a primary factor of personality derived by a factorial analysis of ratings.
- Spearman-Brown prophecy formula** A method of estimating the correlation most likely to be obtained between two longer forms of a test when that between comparable short forms and the relative number of items in the long and the short forms is known.
- standard** An object, an attribute, or a quantity used as a basis with which others may be compared. A reference standard is one chosen for use in a particular investigation; it may or may not be suitable for other purposes.
- standard deviation (S.D.)** A measure of variability found by subtracting each individual number or score in a *frequency distribution* from the mean of the distribution, squaring these differences, summing them, dividing by the number of cases and finding the square root of the result.
- standard error** An estimate, based upon the variability of scores within a *sample*, of the *variability* of a given *statistic* in a *universe*.
- standard score** A score expressed in terms of the number of *standard deviations* by which it exceeds or falls below the *mean* of the group to which the subject belongs.
- standardization** In mental testing, the process of establishing standard procedures and normative values for the comparison and evaluation of the performances of individuals or of groups.
- standardization group** The group of subjects whose performances are used for determining the *norms* or *standards* with which those of others are to be compared.
- statistic** A numerical value obtained by mathematical computation which is taken to represent some property of a *universe* or of a sample of that universe.
- statistics** (1) A group or series of such derived values as "the *statistics*

of population growth"; (2) the science which deals with methods of deriving statistical values.

stereotype An act of behavior or a mode of thinking which tends to take the same form with little regard to variations in the situation which gives rise to it.

sthenic Tending to increased strength or vigor.

structure That aspect of an organized whole which has to do with the position, arrangement, and interdependence of its parts. A "structured situation" is one in which so much of the organizing process has already been carried out that it can be dealt with in only a limited number of ways; an "unstructured situation" is one that permits a well-nigh indefinite number of ways of handling.

subjective Originating in and thus peculiar to an individual; hence not directly susceptible to verification by others.

subnormal Below the normal or usual level.

surface traits *Traits* that can be noted by surface observation without recourse to mathematical treatment.

surgency A general tendency toward cheerfulness and optimism; a care-free state.

sympathetic magic The belief that a symbol and the thing symbolized have some mutual or "sympathetic" effect upon each other, such that by mutilation of an image, for example, the person or thing represented by the image will be mutilated in a similar way.

syndrome A group of signs or symptoms which jointly point to a single condition such as a disease or a chronic ailment of some particular kind.

T

t A *statistic* used in estimating the probability of exceeding a given divergence, which takes account of the difference between the *variance* of a *sample* and that of the *universe* from which it is drawn.

T-score A method of scaling test scores originally proposed by McCall in which the unit of measurement was 0.1 of the *standard deviation* of scores made by a representative sample of twelve-year-olds, and the mean of the twelve-year group was set at 50. In its modern usage, the original scores are first transformed into *standard scores* on the basis of the distribution for any specified group which is being handled. Transformation into T-scores is then made by substituting 50 for the obtained mean and adding to or subtracting from this, 1 point for each tenth of a standard deviation above or below the mean of the group.

tachistoscope An instrument for presenting visual stimuli (such as

objects, words, or letters) for definite short periods of time; used particularly in measuring the *span of visual apprehension*.

test A task or series of tasks given to individuals or to groups with the purpose of ascertaining their relative proficiency as compared to each other or to standards previously set up on the basis of the performance of other similar groups.

tetrachoric r A method of computing the *correlation* between two *continuous variables* that have been artificially constricted into the form of a *dichotomy*.

tetrad Any four homologically situated *variables* chosen from a *correlation matrix*.

tetrad difference The difference between the products of the diagonal values of the correlations of the four variables comprising a *tetrad*. If a single factor is sufficient to account for the intercorrelations, the two cross products will not differ from each other by an amount greater than can be attributed to chance.

Thematic Apperception Test (TAT) A projective device developed by Murray in which the subject is shown a standard series of pictures and asked to make up a short story about each one. The stories are scored in terms of certain fairly objective *criteria* having to do with specific aspects of the themes chosen.

theory The tentative formulation of a rule or principle on the basis of a considerable body of evidence which is, however, not sufficiently crucial to demand universal acceptance of the theory or to elevate it to the rank of a scientific *law*.

thyroid gland One of the glands of internal secretion. It is located on the larynx in the neck and secretes a *hormone* of marked importance for the normal mental and physical development of the individual.

time sampling A method of measuring behavior in terms of the number of short-time samples in a systematically planned set of observations during which a given type of behavior is observed.

tonometer An instrument for producing tones of a given pitch or for measuring the pitch of tones.

tonoscope An instrument by which complex tones are analyzed into their components and the results of the analysis shown visually by the intermittent lighting of a series of perforations on a rotating drum.

trait A distinctive mode of behavior of a more or less permanent nature, arising from the individual's native endowments as modified by his experience.—Warren (1934).

true score The average of an infinite number of comparable scores.

U

universe All the parts or elements which together make up an organized system considered as a whole.

behavioral a universe made up of acts of behavior.

V

V-scope See *manoptoscope*.

validation The process of ascertaining a test's *validity* in terms of its *correlation* with some accepted *criterion* or of the extent to which it differentiates between groups of individuals known to differ with respect to the *trait* which the test purports to measure.

validity (1) In mental measurement the term is defined as "the degree to which a test measures that which it purports to measure" (Otis); (2) in a more general sense a conclusion is said to be *valid* if it is a logical deduction from the premises assumed.

variability The extent of dispersion of the measures in a given *frequency distribution*. The most commonly used measures of variability are the *standard deviation*; the *variance*, which is the square of the standard deviation; the *average deviation*; the *range*; the *probable error*; and the *semi-interquartile range*.

variable A characteristic of an individual or of his behavior which may take any one of a series of quantitative values.

independent One in which the scores are taken as the starting point from which others are to be estimated.

dependent One in which the scores vary more or less exactly in accordance with variations in the scores on the *independent variable*.

variance The square of the *standard deviation*.

analysis of See *analysis of variance*.

visual apprehension, span of The number or range of objects, letters, or words that can be recognized at a single glance well enough to permit later report of what was seen.

Y

year scale See *age scale*.

Z

z The height of the *ordinate* at any specified point on the *abscissa* of a *normal distribution*.

z-function (Fisher's) A series of equally spaced units corresponding to those which would be taken by *r* if its sampling distribution were symmetrical.

Bibliography*

- ABBOTT, A., and TRABUE, M. R. *Exercises in judging English poetry*. New York: Bureau of Publications, Teachers College, Columbia University, 1921.
- ABBOTT, GRACE. *The child and the state*. Vol. II, *The dependent and the delinquent child*. Chicago: University of Chicago Press, 1938.
- ADKINS, DOROTHY C., et al. *Construction and analysis of achievement tests. The development of written and performance tests of achievement for predicting job performance of public personnel*. Washington, D.C.: U.S. Government Printing Office, 1947.
- ALLPORT, GORDON W., and ODBERT, HENRY S. "Trait-names, a psycho-lexical study." *Psychological Monographs*, 1936, Vol. 47, No. 1.
- , and VERNON, PHILIP E. *A study of values. Manual of directions*. Boston: Houghton Mifflin Company, rev. ed., 1931.
- . *Studies in expressive movement*. New York: The Macmillan Company, 1933.
- ALSCHULER, ROSE H., and HATTWICK, LABERTA WEISS. *Painting and personality; a study of young children*. Chicago: University of Chicago Press, 1947, 2 vols.
Profusely illustrated.
- AMERICAN COUNCIL ON EDUCATION PSYCHOLOGICAL EXAMINATION. See Thurstone, L. L. and T. G.
- ANDERSON, H. H. "Domination and integration in the social behavior of young children in an experimental play situation." *Genetic Psychology Monographs*, 1937, 19, 341-408.
- ANDERSON, JOHN E. "The limitations of infant and preschool tests in the measurement of intelligence." *Journal of Psychology*, 1939, 8, 351-379.
- ANDREW, DOROTHY M., and PATERSON, DONALD G. *Manual of directions for the Minnesota vocational test for clerical workers*. New York: The Psychological Corporation, 1941.
- ARMITAGE, STEWART G. "An analysis of certain psychological tests used for the evaluation of brain injury." *Psychological Monographs*, 1946, Vol. 60, No. 1.
- ARRINGTON, RUTH E. "Time-sampling studies of child behavior." *Psychological Monographs*, 1939, Vol. 51, No. 2.
- ARTHUR, GRACE. *A point scale of performance tests*. Vol. I, *Clinical manual*. New York: The Commonwealth Fund, 1930.
- . *A point scale of performance tests*. Vol. II, *The process of standardization*. New York: The Commonwealth Fund, 1933.

* NOTE: In periodical references, the volume numbers are set in italic.

- . *Tutoring as therapy*. New York: The Commonwealth Fund, 1946.
- , and WOODROW, HERBERT. "An absolute intelligence scale: a study in method." *Journal of Applied Psychology*, 1919, 3, 118-137.
- ASHER, E. S. "The inadequacy of current intelligence tests for testing Kentucky mountain children." *Journal of Genetic Psychology*, 1935, 46, 480-486.
- ATKINS, RUTH E. *The measurement of the intelligence of young children by an object-fitting test*. Minneapolis: University of Minnesota Press, 1931.
- BABCOCK, HARRIET. "An experiment in the measurement of mental deterioration." *Archives of Psychology*, No. 117, 1930.
- . *Time and the mind; personal tempo the key to normal and pathological mental conditions*. Cambridge, Mass.: Sci-Art Publishers, 1941.
- This is a revision and extension of the preceding monograph.
- BAIN, ALEXANDER. *The senses and the intellect*. London: J. W. Parker and Son, 1855.
- BAKER, H. J. "A mechanical aptitudes test." *Detroit Educational Bulletin*, 1929, 12, 5-6.
- BARUCH, DOROTHY W. "Doll play in preschool as an aid in understanding the child." *Mental Hygiene*, 1940, 24, 566-577.
- BAYLEY, NANCY. "Mental growth in young children." *Thirty-ninth Yearbook of the National Society for the Study of Education*, II, 11-47. Bloomington, Ill.: Public School Publishing Company, 1940.
- . *The California first year mental scale*. Berkeley: University of California Syllabus Series No. 243, 1933.
- . "The development of motor abilities during the first three years." *Monographs of the Society for Research in Child Development*, 1935, Vol 1, No. 1.
- BECK, SAMUEL J. "Introduction to the Rorschach method: a manual of personality study." *American Orthopsychiatric Association Monographs*, No. 1, 1937.
- BELL, HUGH M. *The adjustment inventory. Manual of directions and norms. Student form*. Stanford University, Calif.: Stanford University Press, 1934. (Adult form, 1937.)
- . *The school inventory; manual of directions and norms*. Stanford University, Calif.: Stanford University Press, 1939.
- BELL, JOHN E. *Projective techniques. A dynamic approach to the study of personality*. New York: Longmans, Green & Co., 1948.
- BENNETT, GEORGE K. *Test of mechanical comprehension*. New York: The Psychological Corporation, 1940.
- BERDIE, RALPH F. "Measurement of adult intelligence by drawings." *Journal of Clinical Psychology*, 1945, 1, 288-295.
- BERNREUTER, ROBERT G. *Manual for the personality inventory*. Stanford University, Calif.: Stanford University Press, 1931.

- BINET, ALFRED. *Psychologie des grands calculateurs et joueurs d'échecs*, 1894.
- . "La mesure en psychologie individuelle." *Revue philosophique*, 1898, 46, 113-123.
- . *L'Étude expérimentale de l'intelligence*. Paris: Ancienne Librairie Schleicher, 1902.
- . "Nouvelles recherches sur la mesure du niveau intellectuel chez les enfants d'école." *L'Année psychologique*, 1911, 17, 145-210.
- , and HENRI, V. "La psychologie individuelle," *L'Année psychologique*, 1896, 2, 411-465.
- , and SIMON, TH. "Application des méthodes nouvelles au diagnostic du niveau intellectuel chez des enfants normaux et anormaux d'hospice et d'école primaire." *L'Année psychologique*, 1905, 11, 245-266.
- . "Méthodes nouvelles pour le diagnostic scientifique des états inférieurs de l'intelligence." *L'Année psychologique*, 1905, 11, 163-190.
- . "Le développement de intelligence chez les enfants." *L'Année psychologique*, 1908, 14, 1-94.
- BINGHAM, W. V. *Aptitudes and aptitude testing*. New York: Harper & Brothers, 1937.
- BOBERTAG, O. "Über Intelligenz prüfungen (nach der Methode von Binet und Simon)." *Zeitschrift für Angewandte Psychologie*, 1911, 5, 105-203.
- BOCHNER, RUTH, and HALPERN, FLORENCE. *The clinical application of the Rorschach test*. New York: Grune and Stratton, 1942.
- BOLTON, T. L. "The growth of memory in school children." *American Journal of Psychology*, 1892, 4, 362-380.
- BORING, E. G. *A history of experimental psychology*. New York: Appleton-Century-Crofts Company, 1929.
- BRACE, D. E. *Motor tests*. New York: A. S. Barnes and Company, 1927.
- BRADWAY, KATHERINE P. "IQ constancy on the revised Stanford-Binet from the preschool to the junior high school level." *Journal of Genetic Psychology*, 1944, 65, 197-217.
- . "An experimental study of factors associated with Stanford-Binet IQ changes from the preschool to the junior high school." *Journal of Genetic Psychology*, 1945, 66, 107-128.
- , and HOFFEDITZ, E. LOUISE. "The basis for the personal constant." *Journal of Educational Psychology*, 1937, 28, 501-513.
- BRAY, CHARLES W. *Psychology and military proficiency*. Princeton, N.J.: Princeton University Press, 1948.
- BROWN, WILLIAM. *The essentials of mental measurement*. London: Cambridge University Press, 1911.
- BUCK, JOHN N. *The H-T-P Test*. (Mimeographed material obtainable from the author.) Lynchburg State Colony, Colony, Virginia: Department of Mental Hygiene and Hospitals. No date.

- BÜHLER, CHARLOTTE, and HETZER, HILDEGARD. *Kleinkindertests. Entwicklungs-tests von 1-6 Lebensjahr*. Leipzig: Barth, 1932.
- , ———, and TUDOR-HART, BEATRIX. *Soziologische und psychologische Studien über das erste Lebensjahr. I, Die ersten sozialen Verhaltensweisen des Kindes*. Jena: Gustav Fischer, 1927.
- , and KELLEY, G. *The World Test. A measure of emotional disturbance*. New York: The Psychological Corporation, 1941.
- BURGESS, ERNEST W., and COTTRELL, LEONARD S. *Predicting success or failure in marriage*. New York: Prentice-Hall, Inc., 1939.
- BURKS, BARBARA STODDARD, JENSEN, DORTHA WILLIAMS, and TERMAN, LEWIS M. *Genetic studies of genius*. Vol. III. *The promise of youth. Follow-up studies of a thousand gifted children*. Stanford University, Calif.: Stanford University Press, 1930.
- BURT, CYRIL. *Mental and scholastic tests*. London: P. S. King and Son, 1921.
- . *The young delinquent*. New York: Appleton-Century-Crofts Company, 1925.
- BURTT, H. E. "Motor concomitants of the association response." *Journal of Experimental Psychology*, 1936, 19, 51-63.
- CANTRIL, HADLEY, and ASSOCIATES. *Gauging public opinion*. Princeton, N.J.: Princeton University Press, 1944.
- CATTELL, JAMES MCKEEN. "Mental tests and measurements." *Mind*, 1890, 15, 373-381.
- CATTELL, PSYCHE. "The Heinis Personal Constant as a substitute for the IQ." *Journal of Educational Psychology*, 1933, 24, 221-228.
- . *The measurement of intelligence of infants and young children*. New York: The Psychological Corporation, 1940.
- CATTELL, RAYMOND B. *Description and measurement of personality*. Yonkers-on-Hudson, N.Y.: World Book Company, 1946.
- CHAILLE, STANFORD E. "Infants: their chronological progress." *New Orleans Medical and Surgical Journal*, 1887, 14, 893-912.
- CHAMPNEY, HORACE. "The measurement of parent behavior." *Child Development*, 1941, 12, 131-166.
- CHESIRE, LEONE, SAFFIR, MILTON, and THURSTONE, L. L. *Computing diagrams for the tetrachoric correlation*. Chicago: University of Chicago Bookstore (distributors), 1933.
- CLARK, RUTH MILLBURN. "A method of administering and evaluating the Thematic Apperception Test in a group situation." *Genetic Psychology Monographs*, 1944, 30, 3-55.
- CONRAD, HERBERT S. *The California behavior inventory for nursery school children*. Berkeley: University of California Press, 1933.
- , FREEMAN, FRANK N., and JONES, HAROLD E. "Differential mental growth." Chap. IX, pp. 164-184, in the *Forty-third Yearbook of the National Society for the Study of Education*, Part I, *Adolescence*. Chicago: University of Chicago Press, 1944.
- CORNELL, E. L., and COXE, W. W. *Cornell-Coxe performance ability scale*.

- Manual of directions.* Yonkers-on-Hudson, N.Y.: World Book Company, 1934.
- COURTIS, S. A. *Courtis standard research tests in arithmetic.* Detroit, Mich.: S. A. Courtis, publisher, 1908.
- COWDERY, K. M. "Measurement of professional attitudes." *Journal of Personnel Research*, 1926, 5, 131-141.
- COX, CATHERINE MORRIS. *Genetic studies of genius.* Vol. II, *The early mental traits of three hundred geniuses.* Stanford University, Calif.: Stanford University Press, 1926.
- CULLEN, WILLIAM. *Synopsis nosologicae medicae*, 1769. American edition entitled *A treatise of the materia medica.* Philadelphia: Mathew Carey, 1808, 2 vols.
- CUNNINGHAM, BESS V. "An experiment in measuring gross motor development of infants and young children." *Journal of Educational Psychology*, 1927, 18, 458-464.
- CUNNINGHAM, KENNETH S. *The measurement of early levels of intelligence.* New York: Bureau of Publications, Teachers College Contributions to Education, No. 259, Columbia University, 1927.
- DARLEY, J. G., et al. *The use of tests in college.* Washington, D.C.: American Council on Education Studies, Series VI, 1947, Vol. 11, No. 9.
- DARWIN, CHARLES. *The origin of species.* London: John Murray, 1859.
- . "A biographical sketch of an infant." *Mind*, 1877, 2, 285-294.
- . *Expression of the emotions in man and animals.* First English edition, 1872; American edition: New York: D. Appleton Company, 1873.
- DAVIDSON, HELEN P. "An experimental study of bright, average, and dull children at the four-year mental level." *Genetic Psychology Monographs*, 1931, 9, 119-289.
- DAVIS, EDITH A. *The development of linguistic skill in twins, singletons with siblings, and only children from five to ten years.* Minneapolis: University of Minnesota Press, University of Minnesota Institute of Child Welfare Monograph, Series No. 14, 1937.
- DAY, ELLA J. "The development of language in twins. I. A comparison of twins and single children." *Child Development*. 1932, 3, 179-199.
- DEPUTY, E. C. *Predicting first grade reading achievement.* New York: Bureau of Publications, Teachers College. Contributions to Education, No. 426, Columbia University, 1930.
- DERI, SUSAN K. *Introduction to the theory and practice of the Szondi Test.* New York: Grune and Stratton, 1948.
- DESPERT, J. LOUISE. "A method for the study of personality reactions in preschool children by means of analysis of their play." *Journal of Psychology*, 1940, 9, 17-29.
- DIVISION OF PSYCHOLOGY, SURGEON GENERAL'S OFFICE. *Army group intelli-*

- gence examinations; *Alpha, Beta, and point scales*. Washington, D.C., 1918.
- DOLL, E. A. *The Vineland social maturity scale. Revised and condensed manual of directions*. Vineland, N.J.: Publication of The Training School Department of Research, Series 1936, No. 3, 1936.
- DOWNEY, JUNE E. "Judgments on the sex of handwriting." *Psychological Review*, 1910, 17, 205-216.
- EBBINGHAUS, H. *Über das Gedächtniss*, 1885. Translation by H. Ruger, *On memory*. New York: Bureau of Publications, Teachers College, Columbia University, 1913.
- EBERT, ELIZABETH H. "A comparison of the original and revised Stanford-Binet Scales." *Journal of Psychology*, 1941, 11, 47-61.
- ELLIS, ALBERT. "The validity of personality questionnaires." *Psychological Bulletin*, 1946, 43, 385-440.
- ESPENSHADE, ANNA. "Motor performance in adolescence, including the study of relationships with measures of physical growth and maturity." *Monographs of the Society for Research in Child Development*, 1940, Vol. 5, No. 1.
- ESQUIROL, JEAN-ETIENNE DOMINIQUE. *Des Maladies mentales considérées sous les rapports médical, hygienique, et médico-légal*. Paris: J. B. Baillière, 1838, 2 vols. and atlas.
- . *Die geisteskrankheiten in beziehung zur medizin und staatsarzneikunde vollständig dargestellt, von E. Esquirol . . . Ins deutsche übertragen von dr. w. Bernard . . .* Berlin: Voss, 1838.
- FISHER, R. A. *Statistical methods for research workers*. Edinburgh and London: Oliver & Boyd, 1936.
- . *The design of experiments*. Edinburgh and London: Oliver & Boyd, 1937.
- , and YATES, F. *Statistical tables for biological, agricultural and medical research*. Edinburgh and London: Oliver & Boyd, 1938.
- FLANAGAN, JOHN CLEMENS. *Factor analysis in the study of personality*. Stanford University, Calif.: Stanford University Press, 1935.
- FRANK, L. K. "Projective methods for the study of personality." *Journal of Psychology*, 1939, 8, 389-413.
- FRANZEN, RAYMOND. "The accomplishment ratio." *Teachers College Record*, 1920, 21, 432-442.
- FREYD, MAX. "A method for the study of vocational interests." *Journal of Applied Psychology*, 1922, 6, 243-254.
- FULLERTON, G. S., and CATTELL, J. McK. *On the perception of small differences*. Philadelphia: University of Pennsylvania Press, Philosophical Series of the Publications of the University of Pennsylvania, No. 2, 1892.
- FURFEY, PAUL HANLY. "A revised scale for measuring developmental age in boys." *Child Development*, 1931, 2, 102-114.
- GALTON, SIR FRANCIS. *Hereditary genius*. London: Macmillan and Company, Ltd., 1869.

- . *English men of science: their nature and nurture*. London: Macmillan and Company, Ltd., 1874.
- . *Inquiries into human faculty and its development*. London: E. P. Dutton and Company, 1883.
- . *Natural inheritance*. London and New York: The Macmillan Company, 1889.
- GARRETT, H. E. *Statistics in psychology and education*. New York: Longmans, Green & Co., 3rd ed., 1947.
- GARRISON, F. H. *An introduction to the history of medicine*. Philadelphia: W. B. Saunders Company, 1929.
- GATES, ARTHUR I. *Gates Reading Readiness Tests*. New York: Bureau of Publications, Teachers College, Columbia University, 1939.
- GESELL, ARNOLD. *Infancy and human growth*. New York: The Macmillan Company, 1928.
- GILBERT, J. A. "Researches on the mental and physical development of school children." *Studies from the Yale Psychological Laboratory*, 1894, 2, 40-100.
- . "Researches upon school children and college students." *University of Iowa Studies in Psychology*, 1897, 1, 1-39.
- GILLILAND, A. R. "A revision and some results with the Moore-Gilliland aggressiveness test." *Journal of Applied Psychology*, 1926, 10, 143-150.
- . "What do introversion-extroversion tests measure?" *Journal of Abnormal and Social Psychology*, 1934, 28, 407-412.
- GLUECK, SHELDON, and GLUECK, ELEANOR. *One thousand juvenile delinquents: their treatment by court and clinic*. Cambridge, Mass.: Harvard University Press, 1934.
- GODDARD, H. H. "Four hundred feeble-minded children classified by the Binet method." *Pedagogical Seminary*, 1910, 17, 387-397.
- . "A measuring scale for intelligence." *The Training School*, 1910, 6, 146-155.
- . "Two thousand normal children measured by the Binet measuring scale of intelligence." *Pedagogical Seminary*, 1911, 18, 232-259.
- . *The Kallikak family*. New York: The Macmillan Company, 1912.
- . *Feeble-mindedness: Its causes and consequences*. New York: The Macmillan Company, 1914.
- . "In defense of the Kallikak study." *Science*, 1942, 95, 574-576.
- GOLDSTEIN, KURT. *The organism: a holistic approach to biology derived from pathological data in man*. New York: American Book Company, 1939.
- , and SCHEERER, MARTIN. "Abstract and concrete behavior. An experimental study with special tests." *Psychological Monographs*, 1941, Vol. 33, No. 2.
- GOODENOUGH, FLORENCE L. "The reading tests of the Stanford Achievement Scale and other variables." *Journal of Educational Psychology*, 1925, 16, 525-531.

- . *The measurement of intelligence by drawing*. Yonkers-on-Hudson, N.Y.: World Book Company, 1926.
- . *The Kuhlmann-Binet tests for children of preschool age: a critical study and evaluation*. Minneapolis: University of Minnesota Press, 1928.
- . "The relation of the intelligence of preschool children to the occupation of their fathers." *American Journal of Psychology*, 1929, 40, 284-294.
- . "Children's drawings." Chapter 14 in *A handbook of child psychology* (edited by Carl Murchison). Worcester, Mass.: Clark University Press, 1931.
- . "A critical note on the use of the term 'reliability' in mental measurement." *Journal of Educational Psychology*, 1936, 27, 173-178.
- . "Studies of the 1937 Revision of the Stanford-Binet Scale. I. Variability of the IQ at successive age-levels." *Journal of Educational Psychology*, 1942, 33, 241-251.
- . "The use of free association in the objective measurement of personality." In *Studies in personality contributed in honor of Lewis M. Terman*. New York: McGraw-Hill Book Company, 1942.
- . *Developmental psychology*. New York: Appleton-Century-Crofts Company, rev. ed., 1945.
- . "Sex differences in judging the sex of handwriting." *Journal of Social Psychology*, 1945, 22, 61-68.
- . "Semantic choice and personality structure." *Science*, 1946, 104, 451-456.
- , and ANDERSON, JOHN E. *Experimental child study*. New York: Appleton-Century-Crofts Company, 1931.
- , and MAURER, KATHARINE M. *The mental growth of children from two to fourteen years; a study of the predictive value of the Minnesota Preschool Scales*. Minneapolis: University of Minnesota Press, 1942.
- , and VAN WAGENEN, M. J. *Minnesota Preschool Scales. Forms A and B*. Minneapolis: Educational Test Bureau, rev. ed., 1940.
- GREEN, H. J., BERMAN, I. R., PATERSON, D. G., and TRABUE, M. R. *A manual of selected occupational ability tests*. Minneapolis: University of Minnesota Press, 1933.
- GREENE, KATHARINE B. "The influence of specialized training on tests of general intelligence." Chapter XXI in *The Twenty-seventh Yearbook of the National Society for the Study of Education*. Bloomington, Ill.: Public School Publishing Company, 1928.
- GREENE, RONALD R. "Studies in pilot selection. II. The ability to perceive and react differentially to configurational changes as related to the piloting of light aircraft." *Psychological Monographs*, 1947, Vol. 61, No. 5.
- GUILFORD, J. P. "An examination of a typical test of introversion-extro-

- version by means of the method of similar reactions." *Journal of Social Psychology*, 1933, 4, 430-443.
- . "Introversion-extroversion." *Psychological Bulletin*, 1934, 31, 331-354.
- . *Psychometric methods*. New York: McGraw-Hill Book Company, 1936.
- . *Inventory of factors S, T, D, C, R. Manual of directions and test forms*. Beverly Hills, Calif.: Sheridan Supply Company, 1939.
- . "Some lessons from aviation psychology." *American Psychologist*, 1948, 3, 3-11. (Presidential address read before the Western Psychological Association, June 19, 1947.)
- , and GUILFORD, R. B. "Personality factors, D, R, T, and A." *Journal of Abnormal and Social Psychology*, 1939, 34, 21-36.
- , and LACEY, JOHN I. (EDITORS). *Printed classification tests*. Washington, D.C.: Government Printing Office, Army Air Forces Aviation Psychology Program, Report No. 5, 1947.
- , and MARTIN, HOWARD G. *The Guilford-Martin Personnel Inventory*. Beverly Hills, Calif.: Sheridan Supply Company, 1943.
- GUTTERIDGE, MARY V. "A study of motor achievements of young children." *Archives of Psychology*, No. 244, 1939.
- GUTTMAN, LOUIS. "A basis for scaling quantitative data." *American Sociological Review*, 1944, 9, 139-150.
- HAGGERTY, M. E., and NASH, H. B. "Mental capacity of children and paternal occupation." *Journal of Educational Psychology*, 1924, 15, 559-572.
- HALL, B. F. "The trial of William Freeman." *American Journal of Insanity*, 1848, 5, 34-60.
- HALVERSON, H. M. "An experimental study of prehension in infants by means of systematic cinema records." *Genetic Psychology Monographs*, 1931, 10, 107-286.
- HARRISON, R. "Validation by the method of 'blind analysis.'" *Character and Personality*, 1940, 9, 134-138.
- HARROWER-ERICKSON, M. R., and STEINER, M. E. *Large scale Rorschach techniques*. Springfield, Ill.: Charles C Thomas, 1945.
- HARTSHORNE, H., and MAY, M. A. *Studies in the nature of character*. Vol. I, *Studies in deceit*. New York: The Macmillan Company, 1928.
- , and MALLER, J. B. *Studies in the nature of character*. Vol. II, *Studies in service and self-control*. New York: The Macmillan Company, 1929.
- , and SHUTTLEWORTH, F. K. *Studies in the nature of character*. Vol. III, *Studies in the organization of character*. New York: The Macmillan Company, 1930.
- HATHAWAY, S. R., and MCKINLEY, J. C. "A multiphasic personality schedule (Minnesota): I. Construction of the schedule." *Journal of Psychology*, 1940, 10, 249-254. (Material and test blanks distributed by The Psychological Corporation, New York.)

- HAWKES, HERBERT E., LINDQUIST, E. F., and MANN, C. R. *The construction and use of achievement examinations*. Boston: Houghton Mifflin Company, 1936.
- HAYES, ELINOR G. "Selecting women for shop work." *Personnel Journal*, 1932, 11, 69-85.
- HEINIS, H. "La loi du developpement mental." *Archives de psychologie*, 1924, No. 74, pp. 97-128.
- . "A personal constant." *Journal of Educational Psychology*, 1926, 17, 163-186.
- HENRY, WILLIAM E. "The thematic apperception technique in the study of culture-personality relations." *Genetic Psychology Monographs*, 1947, 35, 3-135.
- HERRING, JOHN P. *Herring revision of the Binet-Simon tests*. Yonkers-on-Hudson, N.Y.: World Book Company, 1922.
- HILDEN, A. H. "A comparative study of the intelligence quotient and Heinis' personal constant." *Journal of Applied Psychology*, 1933, 17, 355-375.
- . *Table of Heinis' personal constant values*. Minneapolis: Educational Test Bureau, 1933.
- HILDRETH, GERTRUDE. *Bibliography of mental tests and rating scales*. New York: The Psychological Corporation, 2d ed., 1939. *Supplement to above*, 1945.
- , and GRIFFITHS, N. L. *Metropolitan Readiness Tests*. Yonkers-on-Hudson, N.Y.: World Book Company, 1933.
- HOLLINGWORTH, LETA S. *Psychology of subnormal children*. New York: The Macmillan Company, 1920.
- . *Children above 180 IQ, Stanford-Binet. Origin and development*. Yonkers-on-Hudson, N.Y.: World Book Company, 1942.
- HOLZINGER, KARL J., and FREEMAN, FRANK N. "The interpretation of Burt's regression equation." *Journal of Educational Psychology*, 1925, 16, 577-582.
- HULL, CLARK L. *Aptitude testing*. Yonkers-on-Hudson, N.Y.: World Book Company, 1928.
- . *Hypnosis and suggestibility*. New York: Appleton-Century-Crofts Company, 1933.
- HUMM, D. G., and WADSWORTH, G. W., JR. *The Humm-Wadsworth Temperament Scale. Manual of directions*. Los Angeles: The D. G. Humm Personal Service, rev. ed., 1940.
- HUMPHREY, GEORGE, and HUMPHREY, MURIEL. *The wild boy of Aveyron*. New York: Appleton-Century-Crofts Company, 1932.
- . Translation of Itard's reports.
- HUNT, J. McV. (EDITOR). *Personality and the behavior disorders: a handbook based on experimental and clinical research*. New York: The Ronald Press, 1944, 2 vols.
- HUNTER, W. S. "The delayed reaction in animals and children." *Behavior Monographs*, 1913, 2, 1-86.

- HURLOCK, ELIZABETH B. "The value of praise and reproof as incentives for children." *Archives of Psychology*, 1924, Vol. 11, No. 71.
- ITARD, JEAN-MARC GASPARD. See Humphrey, George and Muriel.
- JANET, PIERRE. *The major symptoms of hysteria*. New York: The Macmillan Company, 1907.
- JASPEN, NATHAN. "A note on the age-placement of Binet tests." *Psychological Bulletin*, 1944, 41, 41-42.
- JASTROW, JOSEPH. "Some anthropological and psychological tests on college students—a preliminary survey." *American Journal of Psychology*, 1892, 4, 420-427.
- JOHNSON, G. E. "Contributions to the psychology and pedagogy of feeble-minded children." *Pedagogical Seminary*, 1894, 3, 246-301.
- JOHNSON, WENDELL, and DAVIS, DOROTHY M. "Dextrality quotients of seven-year-olds in terms of hand usage." *Journal of Educational Psychology*, 1937, 28, 346-354.
- , and DUKE, D. "The dextrality quotients of fifty six-year-olds with regard to hand usage." *Journal of Educational Psychology*, 1936, 27, 26-36.
- JORDAN, R. C. "An empirical study of the reliability coefficient." *Journal of Educational Psychology*, 1935, 26, 416-420.
- JUBILÉ DE LA PSYCHOLOGIE SCIENTIFIQUE FRANÇAISE. *Centenaire de Th. Ribot*. Paris: Agen, Imprimerie Moderne, 43 rue Voltaire, 1939.
- KANNER, LEO. *Child psychiatry*. Springfield, Ill. Charles C Thomas, 1935.
- KAPLAN, OSCAR J. (EDITOR). *Encyclopedia of vocational guidance*. New York: Philosophical Library, 1948, 2 vols.
- KATZ, DAVID, and MACLEOD, ROBERT B. "The mandible principle in muscular action." In *Centenaire de Th. Ribot jubilé de la psychologie scientifique française*. Paris: Agen, Imprimerie Moderne, 43 rue Voltaire, 1939.
- KELLEY, TRUMAN L. *Statistical method*. New York: The Macmillan Company, 1923.
- . *Crossroads in the mind of man*. Stanford University, Calif.: Stanford University Press, 1928.
- KELLOGG, W. N., and KELLOGG, L. A. *The ape and the child: a study of environmental influence upon early behavior*. New York: Whittlesley House, 1933.
- KLOPPER, BRUNO, and KELLEY, DOUGLAS MCGLASHAN. *The Rorschach technique: a manual for a projective method of personality diagnosis*. Yonkers-on-Hudson, N.Y.: World Book Company, 1942.
- KNOX, H. A. "A scale based on the work at Ellis Island for estimating mental defect." *Journal of the American Medical Association*, 1914, 62, 741-747.
- KUDER, G. F. "The stability of preference items." *Journal of Social Psychology*, 1939, 10, 41-50.
- . *Kuder Preference Record, Form BB*. Chicago: Science Research Associates, rev. ed., 1942.

- . *Intermediate manual for the Kuder Preference Record*. Chicago: Science Research Associates, 1944.
- KUHLMANN, F. "Binet and Simon's system for measuring the intelligence of children." *Journal of Psycho-Aesthetics*, 1911, 15, 76-92.
- . "A revision of the Binet-Simon system for measuring the intelligence of children." *Journal of Psycho-Aesthetics, Monograph Supplement*, 1912.
- . "The results of repeated mental re-examinations of six hundred thirty-nine feeble-minded over a period of ten years." *Journal of Applied Psychology*, 1921, 5, 195-224.
- . *A handbook of mental tests*. Baltimore: Warwick and York, 1922.
- . *Tests of mental development. A complete scale for individual examination*. Minneapolis: Educational Test Bureau, 1939.
- , and ANDERSON, ROSE G. *Kuhlmann-Anderson Intelligence Tests*. Minneapolis: Educational Test Bureau, 5th ed., 1940.
- LANE, G. GORHAM. "Studies in pilot selection. I. The prediction of success in learning to fly light aircraft." *Psychological Monographs*, 1947, Vol. 61, No. 5.
- LEAHY, ALICE M. *The measurement of urban home environment*. Minneapolis: University of Minnesota Press, 1936.
- LEWIN, KURT. *A dynamic theory of personality. Selected papers*. (Translated by Donald K. Adams and Karl E. Zener.) New York: McGraw-Hill Book Company, 1935.
- . *Principles of topological psychology*. (Translated by Fritz Heider and Grace M. Heider.) New York: McGraw-Hill Book Company, 1936.
- LIKERT, RENSIS. "A technique for the measurement of attitudes." *Archives of Psychology*, No. 140, 1932.
- . "Public opinion polls." *Scientific American*, 1948, 179, 7-11.
- , and QUASHA, W. R. *Minnesota paper form board test. Series AA and BB*. New York: The Psychological Corporation, 1934.
- LINDQUIST, E. F. *Statistical analysis in educational research*. Boston: Houghton Mifflin Company, 1940.
- LOMBROSO, CESARE. *Le Crime, causes et remèdes*, 1899. English translation under the title *The criminal man*. New York: G. P. Putnam's Sons, 1911.
- LÖWENFELD, VIKTOR. *The nature of creative activity*. New York: Harcourt, Brace and Company, 1939.
- LURIA, A. R. *The nature of human conflict, or emotion, conflict, and will; an objective study of disorganization and control of human behavior*. (Translated from the Russian and edited by W. Horsley Gantt.) New York: Liveright Publishing Corporation, 1932.
- MALLER, J. B. "Vital indices and their relation to psychological and social factors." *Human Biology*, 1933, 5, 94-121.

- MARTIN, E. R. "Tests of muscular efficiency." *Physiological Review*, 1921, 1, 454.
- MARTIN, HOWARD G. "Locating the trouble-maker with the Guilford-Martin Personnel Inventory." *Journal of Applied Psychology*, 1944, 28, 461-467.
- . "The construction of the Guilford-Martin inventory of Factors G A M I N." *Journal of Applied Psychology*, 1945, 29, 298-300.
- MAURER, KATHARINE M. *Intellectual status at maturity as a criterion for selecting items on preschool tests*. Minneapolis: University of Minnesota Press, 1946.
- . "Measuring leadership in college women by free association." (Abstract.) *American Psychologist*, 1947, 2, 334.
- MCCALL, W. A. *How to measure in education*. New York: The Macmillan Company, 1922.
- MCCARTY, STELLA AGNES. *Children's Drawings: a study of interests and abilities*. Baltimore: Williams & Wilkins, 1924.
- MCCLOY, C. H. "Measurement of general motor capacity and general motor ability." *Research Quarterly, American Physical Education Association, Supplement*, 1934, 5, 46-61.
- McFARLAND, R. A., and SEITZ, C. P. "Psychosomatic inventory." *Journal of Applied Psychology*, 1938, 22, 327-339.
- MCGRAW, MYRTLE B. "Development of neuromuscular mechanisms as reflected in the crawling and creeping behavior of the human infant." *Journal of Genetic Psychology*, 1941, 58, 83-111.
- . "Neuromuscular maturation of anti-gravity functions as reflected in the development of a sitting posture." *Journal of Genetic Psychology*, 1941, 59, 155-175.
- McKINLEY, J. C., HATHAWAY, S. R., and MEEHL, P. E. "The Minnesota Multiphasic Personality Inventory. VI. The K scale." *Journal of Consulting Psychology*, 1948, 12, 20-31.
- MACKINNON, DONALD W. "The structure of personality." Chapter 1 in *Personality and the behavior disorders: a handbook based on experimental and clinical research*, edited by J. McV. Hunt. New York: The Ronald Press, 1944.
- McNEMAR, QUINN. *The revision of the Stanford-Binet scale*. Boston: Houghton Mifflin Company, 1942.
- . "Opinion-attitude methodology." *Psychological Bulletin*, 1946, 43, 289-374.
- MEEHL, P. E., and HATHAWAY, S. R. "The K factor as a suppressor variable in the Minnesota Multiphasic Personality Inventory." *Journal of Applied Psychology*, 1946, 30, 525-564.
- MEIER, NORMAN C. "Recent research in the psychology of art." Chapter XXVI in *Fortieth Yearbook of the National Society for the Study of Education*. Bloomington, Ill.: Public School Publishing Company, 1941.

- , and SEASHORE, C. E. *Meier-Seashore Art Judgment Test*. Iowa City: University of Iowa Press, 1929.
- MERRILL, MAUD A. *Problems of child delinquency*. Boston: Houghton Mifflin Company, 1947.
- METFESSEL, M., and WARREN, N. D. "Overcompensation by the non-preferred hand in an action-current study of simultaneous movements of the fingers." *Journal of Experimental Psychology*, 1934, 17, 246-256.
- MILES, W. R. "Ocular dominance in adults." *Journal of General Psychology*, 1930, 3, 412-430.
- MOORE, H. T., and GILLILAND, A. R. "The measurement of aggressiveness." *Journal of Applied Psychology*, 1921, 5, 97-118.
- MORENO, J. L. *Who shall survive?* Washington, D.C.: Nervous and Mental Disease Publishing Company, 1934.
- MÜNSTERBERG, H. "Zür individual psychologie." *Centralblatt für Nervenheilkunde und Psychiatrie*, 1891, Vol. 14.
- MURPHY, LOIS BARCLAY. *Social behavior and child personality: an exploratory study of some roots of sympathy*. New York: Columbia University Press, 1937.
- MURRAY, HENRY A., and STAFF OF THE HARVARD PSYCHOLOGICAL CLINIC. *Thematic Apperception Test Manual*. Cambridge, Mass.: Harvard University Press, 1943.
- NATIONAL SOCIETY FOR THE STUDY OF EDUCATION. Forty-fifth Yearbook, Part I. *The measurement of understanding*. Bloomington, Ill.: Public School Publishing Company, 1946.
- . Forty-sixth Yearbook, Part II. *Early childhood education. schools*. Bloomington, Ill.: Public School Publishing Company, 1947.
- . Forty-sixth Yearbook, Part II. *Early childhood education*. Bloomington, Ill.: Public School Publishing Company, 1947.
- . Forty-seventh Yearbook, Part I. *Juvenile delinquency and the schools*. Bloomington, Ill.: Public School Publishing Company, 1948.
- . Forty-seventh Yearbook, Part II. *Reading in the high school and college*. Bloomington, Ill.: Public School Publishing Company, 1948.
- NEWHALL, S. M. "Sex differences in handwriting." *Journal of Applied Psychology*, 1926, 19, 151-161.
- OHIO STATE SCHOOL SURVEY COMMISSION. *Overage and progress in the public schools of Dayton, Ohio*. Dayton: Bureau of Municipal Research, 1914.
- OLSON, WILLARD C. *The measurement of nervous habits in normal children*. Minneapolis: University of Minnesota Press, 1929.
- O'ROURKE, L. J. *O'Rourke Mechanical Aptitude Test. Junior Grade*. New York: The Psychological Corporation, 1937. (Distributors.)
- ORTON, SAMUEL T. "Studies in stuttering. Introduction." *Archives of Neurology and Psychiatry*, 1927, 18, 671-672.

- *OSERETSKY (OSERETZKY) (OZERETZKY), N. I. "Eine metrische Stufenleiter zur Untersuchung der motorische Begabung bei Kindern." *Zeitschrift für Kinderforschung*, 1925, 30, 300-314.
An earlier report of this scale was published in Russia in 1923.
- . "Psychomotorik Methoden zur Untersuchung der Motorik." *Beihefte, Zeitschrift für angewandte Psychologie*, No. 57, 1931.
- OSS ASSESSMENT STAFF. *Assessment of men. Selection of personnel for the Office of Strategic Services*. New York: Rinehart & Company, 1948.
- OTIS, ARTHUR S. *The Otis self-administering tests of mental ability*. Yonkers-on-Hudson, N.Y.: World Book Company, 1922.
- PAGE, HOWARD E. "Detecting psychoneurotic tendencies in army personnel." *Psychological Bulletin*, 1945, 42, 645-658.
- PARSON, B. S. *Lefthandedness*. New York: The Macmillan Company, 1924.
- PARTEN, MILDRED B. "Social participation among preschool children." *Journal of Abnormal and Social Psychology*, 1932, 27, 243-269.
- . "Leadership among preschool children." *Journal of Abnormal and Social Psychology*, 1933, 27, 430-440.
- PATERSON, DONALD G. *The preparation and use of new-type examinations*. Yonkers-on-Hudson, N.Y.: World Book Company, 1925.
- . *Physique and intellect*. New York: Appleton-Century-Crofts Company, 1930.
- PEARSON, KARL. *The grammar of science*. London: Contemporary Science Series, 1892.
- PETERS, CHARLES C., and VAN VOORHIS, WALTER R. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill Book Company, 1940.
- PETERSON, JOSEPH. *Early conceptions and tests of intelligence*. Yonkers-on-Hudson, N.Y.: World Book Company, 1925.
- PINTNER, RUDOLPH. "Nonlanguage Mental Tests." *Journal of Applied Psychology*, 1919, 3, 199-214.
- , EISENSON, JON, and STANTON, MILDRED. *The psychology of the physically handicapped*. New York: Appleton-Century-Crofts Company, 1941.
- , and PATERSON, D. G. *Pintner-Paterson performance test series*. Chicago: C. H. Stoelting Company, 1917.
- . *A scale of performance tests*. New York: Appleton-Century-Crofts Company, 1923.
- PORTEUS, S. D. *Guide to Porteus Maze Test*. Vineland, N.J.: The Training School, Department of Research, Publication No. 25, 1924.
Originally published in 1915, revised 1919, 1924.
- PREYER, W. *Die Seele des Kindes*. Leipzig: Fernau, 1882. English translation by H. W. Brown. *The mind of the child*. Part I, *The senses and the will*; Part II, *The development of the intellect*. New York: D. Appleton Company, 1888, 1889.

* Name is given various spellings in different articles in the literature; the one given first (above) seems most common.

- PRINZHORN, H. *Bildnerei der Geisteskranken: Ein Beitrag zur Psychologie und Psychopathologie der Gestaltung*. Berlin: Springer, 1922.
- QUETELET, L. A. J. *Lettres sur la théorie des probabilités, appliquée aux sciences morales et politiques*, 1846. English translation entitled *Letters on probabilities* by O. G. Downes. London: Layton and Company, 1849.
- RABIN, ALBERT I. "The use of the Wechsler-Bellevue Scales with normal and abnormal persons." *Psychological Bulletin*, 1945, 42, 410-422.
- RAND, GERTRUDE. "A discussion of the quotient method of specifying test results." *Journal of Educational Psychology*, 1925, 16, 599-618.
- RAUBENHEIMER, A. S. *Overstatement test*. Described in Chapter XVII of *Genetic Studies of Genius*, Vol. I, by L. M. Terman, Stanford University, Calif.: Stanford University Press, 1925.
- READ, KATHERINE H. "Significant characteristics of preschool children as located in the Conrad Inventory." *Genetic Psychology Monographs*, 1940, 22, 455-487.
- RIBOT, TH. *La Psychologie anglaise contemporaine*, 1870. English translation (anonymous) entitled *English psychology*. New York: D. Appleton Company, 1874.
- . *L'Hérédité psychologique*, 1873. English translation (anonymous) entitled *Heredity: a psychological study of its phenomena, laws, causes, and consequences*. New York: D. Appleton Company, 1903.
- . *La Psychologie allemande contemporaine*, 1879. English translation by J. M. Baldwin entitled *German psychology of today: the empirical school*. New York: Charles Scribner's Sons, 1886.
- . *Les Maladies de la mémoire*, 1881. English translation by W. H. Smith entitled *Diseases of memory; an essay on the positive psychology*. New York: D. Appleton Company, 1882.
- . *Les Maladies de la volonté*, 1883. English translation by M. M. Snell, entitled *The diseases of the will*. Chicago: Religion of Science Library, 1896.
- . *Les Maladies de la personnalité*, 1885. English translation (anonymous) under title *The diseases of personality*. Chicago: Open Court Publishing Company, 1910.
- . *La Psychologie des sentiments*. Alcan, Paris: 1896. First English translation, 1897. American publication entitled *The psychology of the emotions*. New York: Charles Scribner's Sons, 1911.
- RIDER, PAUL R. *An introduction to modern statistical methods*. New York: John Wiley & Sons, 1939.
- ROGERS, CARL R. *Counseling and psychotherapy*. Boston: Houghton Mifflin Company, 1942.
- RORSCHACH, H. *Psychodiagnostik*. Berne, Switzerland: Ernest Birchen, 1921. (See also Beck, 1937; Klopfer and Kelley, 1942; and others.)
- , and OBERHOLZER, EMIL. "Zür Auswertung des Formdeute versuchs für die Psychoanalyse." *Zeitschrift für die gesamte Neurologie*

- und Psychiatrie*, 1923, 82, 240-274. Translated as "The application of the interpretation of form to psychoanalysis," in *Journal of Nervous and Mental Diseases*, 1924, 60, 225-248, 359-379.
- ROSENZWEIG, S. "The picture-association method and its application in a study of reactions to frustration." *Journal of Personality*, 1945, 14, 3-23.
- ROSS, C. C. *Measurement in today's schools*. New York: Prentice-Hall, Inc., 1941. See especially Chapters 5 and 6.
- RUNDQUIST, EDWARD A. "Inheritance of spontaneous activity in rats." *Journal of Comparative Psychology*, 1933, 16, 415-438.
- , and SLETT, RAYMOND F. *Personality in the depression; a study in the measurement of attitudes*. Minneapolis: University of Minnesota Press, 1936.
- RUST, METTA M. *The effect of resistance on intelligence test scores of young children*. New York: Columbia University Press, 1931.
- SANGREN, P. V. "Comparative validity of primary intelligence tests." *Journal of Applied Psychology*, 1929, 13, 394-412.
- SCHEINFELD, AMRAM. *Women and men*. New York: Harcourt, Brace and Company, 1943.
- SCHERER, MARTIN, ROTHMANN, EVA, and GOLDSTEIN, KURT. "A case of 'Idiot Savant': an experimental study of personality organization." *Psychological Monographs*, 1945, Vol. 58, No. 4.
- SCUPIN, E., and SCUPIN, G. *Bubi's erste Kindheit*. Leipzig: Grieben, 1907.
- SEASHORE, C. E. *The psychology of musical talent*. New York: Silver, Burdett and Co., 1919.
- SEGUIN, EDWARD. *Idiocy: its treatment by the physiological method*. (Reprinted from the original edition of 1866.) New York: Bureau of Publications, Teachers College, Columbia University, 1907.
- SHAKOW, D., and ROSENZWEIG, S. "The use of the Tautophone (verbal summator) as an auditory perceptive test for the study of personality." *Character and Personality*, 1940, 8, 216-226.
- SHAW, CLIFFORD R. *Delinquency areas*. Chicago: University of Chicago Press, 1929.
- SHERMAN, MANDEL, and HENRY, T. R. *Hollow Folk*. New York: The Thomas Y. Crowell Company, 1933.
- SHINN, MILLICENT. *The biography of a baby*. Boston: Houghton Mifflin Company, 1900.
- SHIRLEY, MARY M. *The first two years: a study of twenty-five babies*. Vol. I, *Postural and locomotor development*. Minneapolis: University of Minnesota Press, 1931.
- SKINNER, B. F. "The verbal summator and a method for the study of latent speech." *Journal of Psychology*, 1936, 2, 71-108.
- SNEDECOR, GEORGE W. *Statistical methods applied to experiments in agriculture and biology*. Ames: Iowa State College Press, 4th ed., 1946.
- SOUTH, E. B. *An index of periodical literature on testing, 1921-1936*. New York: The Psychological Corporation, 1937.

- SPEARMAN, C. "'General intelligence' objectively determined and measured." *American Journal of Psychology*, 1904, 15, 201-292.
- . *The abilities of man*. New York: The Macmillan Company, 1927.
- SPENCER, H. *Principles of Psychology*. London: Williams and Norgate, 2d ed., Vol. II, 1872.
- SPRANGER, E. *Lebensformen*. Halle: Niemeyer, 3d ed., 1922. Translation by P. J. W. Pigors entitled *Types of men*. New York: G. E. Stechert Company, 1928.
- STAFF, PERSONNEL RESEARCH SECTION, ADJUTANT GENERAL'S OFFICE. "The Army General Classification Test, with special reference to the construction and standardization of Forms 1a and 1b." *Journal of Educational Psychology*, 1947, 38, 385-420.
- STERN, WILLIAM. *Über Psychologie der individuellen Differenzen. (Ideen zur einer "Differenzellen Psychologie.")* Leipzig: Barth, 1900.
- . *Psychologie der frühen Kindheit, bis zum sechsten Lebensjahre*. Leipzig: Quelle und Meyer, 1914. English translation by A. Barwell, entitled *Psychology of early childhood up to the sixth year of age*. New York: Henry Holt and Company, 1924, rev. ed., 1930.
- STOGDILL, RALPH M. "Attitudes of parents toward parental behavior." *Journal of Abnormal and Social Psychology*, 1934, 29, 293-297.
- STOTT, LELAND H. "Parental attitudes of farm, town, and city parents in relation to certain personality adjustments in their children." *Journal of Social Psychology*, 1940, 17, 325-339.
- . "Parent-adolescent adjustment: its measurement and significance." *Character and Personality*, 1941, 10, 140-150.
- STRAYER, GEORGE D. *Age and grade census of schools and colleges*. Washington, D.C.: Government Printing Office, Report of the United States Bureau of Education, Bulletin 1911, No. 5, 1911.
- STRONG, E. K. "An interest test for personnel managers." *Journal of Personnel Research*, 1926, 5, 194-203.
- . *Vocational interest blanks*. Stanford University, Calif.: Stanford University Press, 1927-1934, revised form, 1938. Vocational Interest blank for women, 1935.
- . *Vocational interests of men and women*. Stanford University, Calif.: Stanford University Press, 1943.
- STUIT, DEWEY B. (EDITOR). *Personnel research and test development in the Bureau of Naval Personnel*. Princeton, N. J.: Princeton University Press, 1948.
- STUTSMAN, RACHEL. *Mental measurement of preschool children with a guide for the administration of the Merrill-Palmer Scale of Mental Tests*. Yonkers-on-Hudson, N.Y.: World Book Company, 1931.
- SYMPOSIUM. "Intelligence and its measurement." *Journal of Educational Psychology*, 1921, 12, 123-147; 195-216.

- TAINÉ, H. *On intelligence*. (Translation by T. D. Haye.) New York: Henry Holt and Company, 1879, 2 vols.
Originally published in France in 1870.
- TAUSSIG, F. W. *Principles of Economics*. New York: The Macmillan Company, Vol. II, 1920.
- TEMPLE, RITA, and AMEN, ELIZABETH W. "A study of anxiety in young children by means of a projective technique." *Genetic Psychology Monographs*, 1944, 30, 59-114.
- TERMAN, L. M. "Genius and stupidity: a study of some of the intellectual processes of seven 'bright' and seven 'stupid' boys." *Pedagogical Seminary*, 1906, 13, 307-373.
- . *The measurement of intelligence*. Boston: Houghton Mifflin Company, 1916.
- . *The intelligence of school children*. Boston: Houghton Mifflin Company, 1919.
- . *The Terman group test of mental ability*. Yonkers-on-Hudson, N.Y.: World Book Company, 1920.
- . "The mental test as a psychological method." *Psychological Review*, 1924, 31, 93-117.
- . *Psychological factors in marital happiness*. New York: McGraw-Hill Book Company, 1938.
- , and ASSOCIATES. *Genetic studies of genius*. Vol I, *Mental and physical traits of a thousand gifted children*. Stanford University, Calif.: Stanford University Press, 1925.
- , and BUTTENWIESER, PAUL. "Personality factors in marital compatibility." *Journal of Social Psychology*, 1935, 6, 143-171; 267-289.
- , and CHILDS, H. G. "Tentative revision and extension of the Binet-Simon measuring scale of intelligence." *Journal of Educational Psychology*, 1912, 3, 61 ff.; 133 ff.; 198 ff.; 277 ff.
- , KOHS, S. C., CHAMBERLAIN, MARY B., ANDERSON, MAYME, and HENRY, BESS. "The vocabulary test as a measure of intelligence." *Journal of Educational Psychology*, 1918, 9, 452-466.
- , and MERRILL, MAUD. *Measuring intelligence: a guide to the administration of the new revised Stanford-Binet tests of intelligence*. Boston: Houghton Mifflin Company, 1937.
- , and MCNEMAR, QUINN. *The Terman-McNemar test of mental ability*. Yonkers-on-Hudson, N.Y.: World Book Company, 1941.
- , MILES, CATHERINE COX, and ASSOCIATES. *Sex and personality: studies in masculinity and femininity*. New York: McGraw-Hill Book Company, 1936.
- , and ODEN, MELITA H. *Genetic studies of genius*. Vol. IV, *The gifted child grows up*. Stanford University, Calif.: Stanford University Press, 1947.
- THOMSON, GODFREY H. "A formula to correct for the effect of errors of

- measurement on the correlation of initial values with gains." *Journal of Experimental Psychology*, 1924, 7, 321-324.
- . "An alternative formula for the true correlation of initial values with gains." *Journal of Experimental Psychology*, 1925, 8, 323-324.
- THORNDIKE, E. L. *The elimination of pupils from school*. Washington, D.C.: Government Printing Office, 1908.
- . "The measurement of achievement in drawing." *Teachers College Record*, 1913, 14, 1-38.
- . *The teachers' word book*. New York: Bureau of Publications, Teachers College, Columbia University, 1921.
- . *The measurement of intelligence*. New York: Bureau of Publications, Teachers College, Columbia University, 1926.
- , WOODYARD, ELLA, and LORGE, IRVING. *Intelligence tests: revised*. New York: Institute of Educational Research, Teachers College, Columbia University, 1935.
- THURSTONE, L. L. "A method of scaling psychological and educational tests." *Journal of Educational Psychology*, 1925, 16, 433-451.
- . "The absolute zero in intelligence measurement." *Psychological Review*, 1928, 35, 175-197.
- . "An experimental study of nationality preferences." *Journal of General Psychology*, 1928, 1, 405-425.
- . "A multiple factor study of vocational interests." *Personnel Journal*, 1931, 10, 198-205.
- . *The vectors of mind: Multiple factor analysis for the isolation of primary traits*. Chicago: University of Chicago Press, 1935.
- . *Primary mental abilities*. Psychometric Monographs, No. 1, 1938.
- . "Current issues in factor analysis." *Psychological Bulletin*, 1940, 37, 189-236.
- . *Multiple factor analysis: a development and expansion of "The vectors of mind."* Chicago: University of Chicago Press, 1947.
- , and CHAVE, E. J. *The measurement of attitude*. Chicago: University of Chicago Press, 1929.
- , and THURSTONE, THELMA GWYNN. *American Council on Education Psychological Examination*. Prepared annually. Published by The American Council on Education, Washington, D.C.
- . *Tests of primary mental abilities for ages 5 and 6. Examiner's manual and test record blanks*. Chicago: Science Research Associates, 1946.
- TIEDEMANN, D. *Beobachtungen über die Entwicklung der Seelen Fähigkeiten bei Kindern*, 1787. English translation by Carl Murchison and Susan Langer under title of "Tiedemann's observations on the development of the mental faculties of children," in *Pedagogical Seminary and Journal of Genetic Psychology*, 1927, 34, 205-230.
- TIFFIN, JOSEPH. *Industrial psychology*. New York: Prentice-Hall, Inc., 1942, rev. ed., 1947.

- TIPPETT, L. H. *The methods of statistics*. London: Williams and Norgate, 2d ed., 1937.
- TOOPS, HERBERT A., and ASSOCIATES. *Ohio State University Psychological Test*. Columbus, Ohio: Committee on Technical Research, Ohio State University (revised annually).
- TRAVIS, LEE EDWARD. *Speech pathology*. New York: Appleton-Century-Crofts Company, 1931.
- TREDGOLD, A. F. *Mental deficiency. (Amentia.)* New York: William Wood and Company, 1908.
- TYRON, ROBERT C. "Individual differences." Chapter 13 in *Comparative psychology*, edited by F. A. Moss. New York: Prentice-Hall, Inc., 1934.
- TYLER, LEONA E. *The psychology of human differences*. New York: Appleton-Century-Crofts Company, 1947.
- VERMEYLEN, G. "Les debiles mentaux (Étude experimentale et clinique)." Paris: *Bulletin de l'institute general psychologie*, No. 4-6, 1922.
- VOELKER, P. F. *The function of ideals and attitudes in social education*. New York: Bureau of Publications, Teachers College Contributions to Education, No. 112, Columbia University, 1921.
- VYGOTSKI, L. S. "Thought in schizophrenia." *Archives of Neurology and Psychiatry*, 1934, 31, 1063-1077.
- WALLIN, J. E. W. *Problems of subnormality*. Yonkers-on-Hudson, N.Y.: World Book Company, 1921.
- WANG, C. K. A. *An annotated bibliography of tests and scales*. Peiping, China: Catholic University Press, Vol. I, 1940. (In English.)
- WARREN, HOWARD C. (EDITOR). *Dictionary of psychology*. Boston: Houghton Mifflin Company, 1934.
- WECHSLER, D. *The measurement of adult intelligence*. Baltimore: Williams and Wilkins, 1st ed., 1939, rev. ed., 1944.
- WELLS, F. L. *The revised Alpha examination. Forms 5 and 7*. New York: The Psychological Corporation, 1932-1933.
- WHIPPLE, G. M. *Manual of mental and physical tests*. Baltimore: Warwick and York, 1st ed., 1910; rev. and enl. ed., 2 vols., 1919 and 1921.
- WILLIAMS, J. HAROLD. "The Whittier Scale for grading home conditions." *Journal of Delinquency*, 1916, 1, 271-286.
- WISSLER, C. L. "The correlation of mental and physical tests." *Psychological Review Monograph Supplements*, 1901, Vol. 3, No. 6.
- WITMER, HELEN LELAND. *Psychiatric clinics for children*. New York: The Commonwealth Fund, 1940.
- . *Psychiatric interviews with children*. New York: The Commonwealth Fund, 1946.
- WITMER, LIGHTNER. *The special class for backward children*. Philadelphia: The Psychological Clinic Press, 1911.
- WOLF, THETA HOLMES. *The effect of praise and competition on the persisting behavior of kindergarten children*. Minneapolis: University of Minnesota Press, 1938.

- WOLFF, WERNER. *The expression of personality*. New York: Harper & Brothers, 1943.
- . *The personality of the preschool child*. New York: Grune and Stratton, 1946.
- . *Diagrams of the unconscious. Handwriting and personality analysis*. New York: Grune and Stratton, 1948.
- WOODRUFF, ASAHIEL DAVIS. *A study of the directive factors in individual behavior*. Privately printed. Distributed by the University of Chicago Libraries, Chicago, 1941.
- WUNDT, WILLIAM. *Grundriss der Psychologie*. Leipzig: William Engelmann, 1896.
- WYMAN, JENNIE B. "Tests of intellectual, social, and activity interests." Chapter XVI, pp. 455-483, in *Genetic studies of genius*. Vol. I, *Mental and physical traits of a thousand gifted children*, by L. M. Terman, et al., Stanford University, Calif.: Stanford University Press, 1925.
- YERKES, ROBERT M. (EDITOR). *Psychological examining in the United States Army*. Washington, D.C.: Memoirs of the National Academy of Sciences, 1921, Vol. 15.
- , BRIDGES, J. W., and HARDWICK, R. S. *A point scale for measuring mental ability*. Baltimore: Warwick and York, 1915.
- YOAKUM, CLARENCE S., and YERKES, ROBERT M. *Army mental tests*. New York: Henry Holt and Company, 1920.
- YOUNG, KIMBALL. *Personality and problems of adjustment*. New York. Appleton-Century-Crofts Company, 1940.
- ZIEVE, LESLIE. "Note on the correlation of initial scores with gains." *Journal of Educational Psychology*, 1940, 31, 391-394.

Indexes

Author Index

A

Abbott, A., 348, 571
 Abbott, G., 12, 571
 Adkins, D. C., 134, 571
 Allport, G. W., 105, 292, 376, 411, 571
 Alschuler, R. H., 426f., 571
 Amen, E. W., 430, 589
 Anderson, J. E., 66, 173, 571, 578
 Andrew, D. M., 450, 571
 Aristotle, 28, 37f.
 Arrington, R. E., 393, 571
 Arthur, G., 69, 201f., 309, 314, 505, 571f.
 Atkins, R. E., 312, 572

B

Babcock, H., 530f., 572
 Backus, F. P., 9
 Bain, A., 21, 572
 Baker, H. J., 475, 476, 572
 Baruch, D. W., 423, 572
 Bateson, W., 53f.
 Bayley, N., 103ff., 175, 310, 362f., 371, 373, 572
 Beck, S. J., 432, 572
 Bell, C., 29
 Bell, H. M., 410, 465, 572
 Bell, J. E., 417, 437, 572
 Bennett, G. K., 475, 572
 Berdie, R. F., 425, 572
 Berkeley, G., 21
 Berman, I. R., 478, 578
 Bernard, W., 5, 576
 Bernreuter, R. G., 410, 501, 572
 Binet, A., 11, 13, 22, 23, 29, 35f., 43ff., 59, 90, 140, 150, 211, 293, 353, 417ff., 432, 521, 573
 Bingham, W. V., 478, 573
 Blackstone, W., 12
 Blakeman, J., 262
 Boas, F., 42
 Bobertag, O., 59, 573
 Bochner, R., 485, 573
 Bolton, T. L., 42, 573
 Boring, E. G., 97ff., 573
 Brace, D. E., 365, 573
 Bradway, K. P., 103, 168, 204, 208, 312, 573
 Braid, J., 22
 Bray, C. W., 501, 573

Bridges, J. W., 59, 211, 592
 Broca, P., 369
 Brown, W., 85, 223, 573
 Buck, J. N., 425, 573
 Bühler, C., 310, 439, 574
 Burgess, E. W., 402, 460, 574
 Burt, C., 253, 523, 525, 574
 Burtt, H. E., 428f., 574
 Buttenwieser, P., 280, 589

C

Cantril, H., 120, 127f., 385f., 574
 Cattell, J. Mc., 13, 29ff., 40f., 42f., 70, 144, 574, 576
 Cattell, P., 207, 310, 312, 574
 Cattell, R. B., 292, 411ff., 574
 Chaille, S. E., 50f., 574
 Champney, H., 405, 465, 574
 Charcot, J. M., 22
 Chesire, L., 266, 574
 Childs, H. G., 59, 589
 Clark, R. M., 430, 574
 Cofer, C. N., 498
 Combe, G., 38
 Conrad, H. S., 174, 406, 574
 Cornell, E. L., 69, 575
 Cottrell, L. S., 402, 460, 574
 Courtis, S. A., 67, 575
 Cowdery, K. M., 76, 575
 Cox, C. M. (C. C. Miles), 511ff., 517, 575
 Coxe, W. W., 69, 575
 Cullen, W., 5, 28, 575
 Cunningham, B. V., 364, 575
 Cunningham, K. S., 152, 154, 575

D

Darley, J. G., 458, 575
 Darwin, C., 23, 24, 31, 575
 Darwin, E., 24
 Davidson, H. P., 323, 575
 Davis, E. A., 143, 575
 Day, E. J., 143, 575
 Deputy, E. C., 323, 575
 Deri, S. K., 438, 575
 Despert, J. L., 423, 575
 DeVries, H., 53
 Doll, E. A., 405, 576
 Downey, J. E., 436, 576

E

- Ebbinghaus, H., 45, 576
 Ebert, E. H., 103, 576
 Eisensohn, J., 529, 585
 Elderton, W. P., 85
 Ellis, A., 79, 406, 408, 501, 539, 576
 Espenschied, A., 365, 576
 Esquirol, J. E. D., 3 ff., 22, 162, 576

F

- Fechner, G. T., 13, 417
 Fisher, R. A., 85, 218 ff., 232 ff., 236, 237, 243, 244, 246, 267, 270, 272, 274, 283, 284, 285, 535, 576
 Flanagan, J. C., 410, 576
 Frank, L. K., 81 f., 416, 576
 Franzen, R., 71, 333, 576
 Freeman, F. N., 174, 253, 574, 580
 Freud, S., 22, 427
 Freyd, M., 76, 576
 Fullerton, G. S., 70, 144, 576
 Furfey, P. H., 402, 576

G

- Gall, F. J., 38
 Gallup, G., 118, 385
 Galton, F., 13, 24 ff., 31, 40 f., 42 f., 54, 85, 576 f.
 Garrett, H. E., 262, 577
 Garrison, F. H., 27, 577
 Gates, A. I., 324, 577
 Gauss, C. F., 24, 165
 Gesell, A., 32, 308, 309, 361 ff., 577
 Gilbert, J. A., 42, 577
 Gilliland, A. R., 402, 577, 584
 Glueck, E., 489 f., 577
 Glueck, S., 489 f., 577
 Goddard, H. H., 51 ff., 60, 62, 532, 577
 Goldstein, K., 352 ff., 531 f., 577, 587
 Goodenough, F. L., 66, 84, 103, 114, 127, 169, 173, 199, 214, 298, 312, 314, 425, 428, 436, 496, 520, 522, 523, 526, 577 f.
 Green, H. J., 478, 578
 Greene, K. B., 167, 578
 Griffiths, N. L., 324, 580
 Guggenbühl, L., 8
 Guilford, J. P., 130, 144, 219, 228, 236, 284, 410, 443, 450, 480, 498, 499 f., 538, 578 f.
 Guttman, L., 130 f., 376, 579

H

- Haggerty, M. E., 114, 579
 Hall, B. F., 10, 579
 Hall, G. S., 31 ff.

- Haller, A. von, 28
 Halpern, F., 485, 573
 Halverson, H. M., 363 f., 579
 Hardwick, R. S., 59, 211, 592
 Harrison, R., 432, 579
 Harrower-Erickson, M. R., 436, 579
 Hartshorne, H., 81, 403, 579
 Hathaway, S. R., 408 ff., 579, 583
 Hattwick, L. W., 426 f., 571
 Hawkes, H. E., 332, 580
 Hayes, E. G., 478, 580
 Heinis, H., 204 ff., 580
 Henri, V., 43 f., 573
 Henry, T. R., 529, 587
 Henry, W. E., 431, 580
 Herring, J. P., 315, 580
 Hilden, A. H., 207, 580
 Hildreth, G., 89, 324, 325, 346, 580
 Hoffeditz, E. L., 204, 208, 573
 Hollingworth, L. S., 391, 521, 580
 Holzinger, K. J., 85, 253, 580
 Howe, S. G., 8 ff.
 Hull, C. L., 218, 402, 580
 Hume, D., 21
 Humm, D. G., 410, 479, 580
 Humphrey, G., 7, 580
 Humphrey, M., 7, 580
 Hunt, J. McV., 498, 580
 Hunter, W. S., 505, 580
 Hurlock, E. B., 142, 581

I

- Itard, J.-M. G., 6 ff., 22

J

- Janet, P., 22, 581
 Jaspens, N., 150, 581
 Jastrow, J., 41, 581
 Johnson, G. E., 9, 581
 Johnson, W., 74, 370, 581
 Jones, H. E., 174, 574
 Jordan, R. C., 214, 581
 Jung, C. G., 427

K

- Kanner, L., 519 f., 581
 Kaplan, O. J., 449, 478, 581
 Katz, D., 368, 581
 Kelley, D. M., 433, 581
 Kelley, T. L., 85, 166, 180, 181, 187, 191, 193, 214, 223, 256, 260, 262, 284, 290, 535, 581
 Kellogg, L. A., 506, 581
 Kellogg, W. N., 506, 581
 Klopfer, B., 433, 581
 Knox, H. A., 69, 581

Kohs, S. C., 4, 589
 Kuder, G. F., 447 f., 450, 478, 581 f.
 Kuhlmann, F., 32, 59, 63, 140, 150, 161,
 205 f., 312, 582

L

Lacey, J. L., 443, 498, 579
 Laplace, P. S. de, 24, 165
 Leahy, A. M., 465, 488, 582
 Levy, D., 432
 Lewin, K., 292 f., 522 f., 538 f., 582
 Likert, R., 380 f., 387 f., 450, 474, 476, 582
 Lindquist, E. F., 219, 236, 238, 267 f., 274,
 285, 332, 580, 582
 Linné, C. von (Linnaeus), 5, 27 f.
 Locke, J., 21, 31
 Lowenfeld, V., 346 f., 582
 Lombroso, C., 39, 582

M

MacKinnon, D. W., 390, 583
 Macmeeken, A., 245
 Magendie, F., 28
 Maller, J. B., 117, 403, 579, 582
 Mann, C. R., 332, 580
 Martin, E. R., 367 f., 583
 Martin, H. G., 450, 480, 583
 Maurer, K. M., 103, 168, 199, 298, 312, 428,
 496, 578, 583
 May, M. A., 81, 403, 579
 McCall, W. A., 196 f., 583
 McCarty, S. A., 348, 583
 McCloy, D. H., 365, 583
 McFarland, R. A., 501, 583
 McGraw, M. B., 362, 583
 McKinley, J. C., 408 ff., 579, 583
 McLeod, R. B., 368, 581
 McNemar, Q., 66, 169, 199, 228, 318, 388 f.,
 447, 537 f., 583, 589
 Meehl, P. E., 409, 583
 Meier, N. C., 347 f., 583 f.
 Mendel, G., 53 f.
 Merrill, M. A., 66, 114, 170, 199, 309, 490 f.,
 584, 589
 Metfessel, M., 75, 584
 Miles, C. C., 402, 526, 589
 Miles, K. A., 464
 Miles, W. R., 75, 584
 Moore, H. T., 402, 584
 Moreano, J. L., 398 ff., 459, 584
 Munsterberg, H., 43, 584
 Murphy, L. B., 394 f., 584
 Murray, H. A., 428 f., 584

N

Nash, H. B., 114, 579

Neilon, P. A., 364
 Newhall, S. M., 436, 584

O

Oberholzer, E., 432, 586
 O'Conner, J., 478
 Odbert, H. S., 105, 292, 411, 571
 Oden, M. H., 255, 349, 589
 Olson, W. C., 395 f., 584
 O'Rourke, L. J., 450, 475, 477, 584
 Orton, S., 354, 369, 584
 Oseretzky, N. I., 74, 364 f., 585
 Otis, A. S., 67, 106, 211, 320, 494, 585

P

Page, H. E., 501, 585
 Parson, B. S., 75, 585
 Parten, M. B., 393 ff., 585
 Paterson, D. G., 37, 69, 134, 209, 332, 571,
 578, 585
 Pearson, K., 25 ff., 31, 85, 236, 535, 585
 Peters, C. C., 191, 236, 265, 585
 Peterson, J., 59, 585
 Pinel, P., 6, 22
 Pintner, R., 69, 209, 316, 468, 529, 585
 Porteus, S. D., 504, 585
 Preyer, W., 31, 585
 Prinzhorn, H., 424, 586

Q

Quasha, W. R., 450, 474, 582
 Quetelet, A., 24 f., 586

R

Rand, G., 176, 586
 Raubenheimer, A. H., 403 f., 586
 Read, K. H., 406, 586
 Reed, W., 510
 Ribot, T., 22 f., 418, 586
 Richet, C., 22
 Rider, P. R., 274, 586
 Rogers, C. R., 460, 586
 Rorschach, H., 83 f., 432 ff., 586 f.
 Rosenzweig, S., 438 f., 587
 Ross, C. C., 134, 332, 587
 Rothman, E., 352 ff., 587
 Rundquist, E. A., 381, 407 f., 506, 586
 Rust, Metta M., 166, 587

S

Saffir, M., 266, 574
 Sangren, P. V., 117, 587
 Sauvages, F. B. de, 5
 Scheerer, M., 352 ff., 531 f., 577, 587
 Scheinfeld, A., 526, 587
 Scupin, E., 31, 587

Scupin, G., 31, 587
 Seashore, C. E., 344f., 584, 587
 Seguin, E., 6ff., 22, 587
 Seitz, C. P., 501, 583
 Shakow D., 437, 587
 Shannon, J. R., 388
 Shaw, C. R., 489, 587
 Sherman, M., 529, 587
 Shinn, M., 31, 587
 Shirley, M. M., 362, 364, 373, 587
 Simon, T., 46ff., 140, 573
 Skinner, B. F., 437, 587
 Sletto, R. F., 381, 407f., 587
 Snedecor, G. W., 219, 246, 249, 274, 587
 South, E. B., 90, 587
 Spearman, C., 85, 223, 226ff., 263, 286ff.,
 338, 500, 517, 537, 588
 Spencer, H., 23, 24, 31, 588
 Spurzheim, G., 38
 Stanton, M., 529, 585
 Stern, W., 31, 45, 63, 140, 161, 588
 Stogdill, R. M., 464, 588
 Stott, L. H., 465, 588
 Stout, G. F., 286
 Strayer, G. D., 17, 588
 Strong, E. K., 76, 280, 447f., 478, 515, 526,
 588
 "Student," 85, 236, 243
 Stuit, D. B., 501, 588
 Stutsman, R., 169, 309, 588

T

Taine, H., 21f., 589
 Taussig, F. W., 115ff., 589
 Temple, R., 430, 589
 Terman, L. M., 4, 32, 59, 62ff., 91f., 106,
 140, 150, 161ff., 170, 199, 245, 255,
 280, 318, 349, 402, 447, 460, 511ff.,
 526, 532, 589
 Thomson, G., 172f., 535, 590
 Thorndike, E. L., 14, 60, 70, 85, 142, 144,
 152, 154, 160, 287ff., 318, 348, 535,
 590
 Thurstone, L. L., 81, 85, 112, 146, 225ff.,
 266, 284, 290f., 314, 378, 381, 448,
 533, 535, 574, 590
 Thurstone, T. G., 229, 314, 590
 Tiedemann, D., 31, 590
 Tiffin, J., 448f., 478, 479, 480, 590

Tippett, L. H., 114, 591
 Trabue, M. R., 348, 571, 578
 Travis, L. E., 369, 591
 Tredgold, A. F., 352, 520, 591
 Tryon, R. C., 506, 591
 Tyler, L. E., 528, 591

V

Van Voorhis, W. R., 191, 236, 265, 585
 Van Wagenen, M. J., 208, 298, 312, 578
 Vermeylen, G., 204f., 208, 591
 Vernon, P. E., 376, 571
 Voelker, P. F., 81, 591

W

Wadsworth, G. W., Jr., 410, 479, 580
 Walker, M., 401
 Wallin, J. E. W., 8, 591
 Wang, C. K. A., 89, 591
 Ward, J., 286
 Warren, H. C., 100f., 344, 377, 390, 591
 Warren, N. D., 75, 584
 Weber, E. H., 13
 Wechsler, D., 309, 316, 320, 403, 447, 523,
 591
 Wells, F. L., 318, 320, 447, 591
 Wenger, M. A., 501
 Whipple, G. M., 365, 591
 Williams, J. H., 488, 591
 Wissler, C., 41, 591
 Witmer, H. L., 466, 591
 Witmer, L., 18f., 591
 Wolff, W., 84, 99, 425f., 437, 438, 592
 Woodrow, H., 201, 572
 Woodruff, A. D., 375f., 592
 Woodworth, R. S., 77f., 495, 534
 Wundt, W., 13, 21, 29ff., 40, 417, 592
 Wyman, J. B., 84, 592

Y

Yates, F., 236, 244, 576
 Yerkes, R. M., 59ff., 67, 211, 495, 592
 Yoakum, C. S., 495, 592
 Young, K., 390, 592
 Yule, G. U., 85, 535

Z

Zieve, L., 172, 592

Subject Index

A

- Ability, distinguished from behavioral tendency, 131
- Absolute scaling, 141 ff., 145
 - definition of, 154, 543
 - relation to year scales, 178
- Accomplishment quotient (A Q), 254
 - 333 ff.
 - definition of, 543
 - sources of error, 334 f.
- Accomplishment ratio (A.R.). (*see* Accomplishment quotient)
- Achievement, compared with intelligence, 333 ff.
 - defined, 543
 - distinguished from aptitude, 325
 - educational, tests of, 322 ff.
- Action current, 74 f., 543
- Age, basal, 50 ff., 59 ff., 151, 545
 - chronological, 64, 543
 - basis for scaling, 150 f.
 - conceptional, 543
 - developmental, 402, 549
 - educational, 550
 - mental (*see* Mental age)
- Age scale (*see* Year scale)
- Aggressiveness, tests of, 402, 427
- Air forces, testing in, 498 ff.
- Altitude of intellect, 152, 160, 543
- American Psychological Association, 13, 67
- Analogies test, 544
 - errors in administering, 93
- Analysis of variance, 219, 221, 271 ff., 292, 544
 - cautions, 281 f.
 - examples, 274 ff., 279 f.
 - requirements, 272
 - as test of linear regression, 262 f.
- Animal intelligence, tests of, 505 f.
- Appropriateness, of test for subject, 125, 146 ff.
- Aptitude tests, educational, 322 f.
 - types of, 72, 325
 - vocational, 75 f., 442 ff., 544
- Arithmetic tests, diagnostic, 357 f.
 - "readiness" tests, 324
 - types of, 71

- Army Alpha Test, 67, 160, 495, 533
 - Wells Revision, 318 ff., 447
- Army Beta Test, 67, 495, 533
- Army General Classification Test, 496, 502
- Array, in correlation surface, 259, 544
- Art talent, 346 ff.
- Arthur Point Scale, 69, 201 ff., 314 f., 317
- Artifact, 190, 544
- Association, free (*see* Free association)
- Assumptions, implicit, 93, 177 ff.
 - of particular problems, 193 f.
- Atkins Object Fitting Test, 312
- Attenuation, correction for, 216 f., 256
 - defined, 545
- Attitude, changes in, 382
 - dimensions of, 378
 - tests of, 81, 127, 377 ff.
 - applications, 382
 - for children, 465
 - construction of, 378 ff.
 - definition of, 545
- Autonomic balance, tests of, 501
- Average, 25, 545
 - choice of, 177

B

- Behavioral universe, distinguished from trait, 99
 - sampled by tests at different ages, 104
- Bernreuter Personality Schedule, 410
- Bias, in criterion, 129 ff.
 - definition of, 545
 - on part of general public, 121 f.
 - on part of mental examiners, 121 f.
 - in public opinion polls, 118 ff., 386 ff.
 - in sampling, 64, 106, 110 f., 112
 - in selection of test items, 125
- Bibliographies, of tests and rating scales, 89 f.
- Binet-Simon tests, 1905 scale, 46 ff.
 - 1908 scale, 49 f.
 - 1911 scale, 51 f.
 - translations and revisions, 51 ff.
- Binet's study of his two daughters, 418 ff.
- Birth fantasy, 426
- Blind analysis, 84, 431 f., 546
- Block building test, 104
- Blood pressure, changes in, 402

Body sway, test of suggestibility, 402
 Boundaries, of behavioral universes, 119f.
 between tension systems, 293
 Brain, localization of functions, 39
 Brain damage, signs of, 352, 498, 505,
 529f.

Buhler Baby-Tests, 310

C

c factor in mental organization, 287
 CAVD test, 152, 160
 California First Year Mental Scale, correlation with Stanford-Binet, 103,
 310f.
 Case study, in clinical practice, 459
 nontypical character of, 466
 subjective factors in, 440f.
 use of tests in, 342f.
 as validating data, 432, 439
 Cattell's Infant Intelligence Scale, 310 f.,
 312
 Cause, 546
 not to be inferred from correlation,
 251f.
 "Ceiling," of test, 147, 546
 Census, U.S., schooling as reported in,
 442
 use in designing samples, 115f.
 Cerebral localization, 39
 birth palsy, 122
 dominance, 74f.
 Chance, definition of, 222
 Cheating, tests of, 81
 Child guidance clinics, basic program, 465
 classes of, 461
 success of treatment, 466
 tests used, 462f.
 Children's drawings (*see* Drawings)
 Chi-square (χ^2), definition, 546
 formula, 236
 sampling distribution, table of, 237
 "Choose your child" test, 464
 Clerical Abilities Test (Minnesota), 450
 Clinical testing, 459ff.
 Coaching, effect of, 167
 indications of, 156
 Coefficient of Intelligence (CI), 211, 546
 College students, mental tests for, 41,
 316f.
 Combat fatigue, 497
 Completion tests, 134
 Compulsory school attendance laws, 14ff.
 Conduct, measurement of, 81
 Confidence, level of (*see* Significance, level
 of)
 Conrad Behavior Inventory, 406

Consistency, internal, 554
 as criterion for item selection, 130f.
 of test results, 79
 Continuous measures, 99, 547
 Cooperation, methods of securing, 297ff.,
 303ff.
 Copying, square vs. diamond, 44
 Cornell-Coxe tests, 69
 Correlation, biserial η , 265
 biserial r , 130, 135f., 264f., 545
 biserial r from widespread classes, 265f.
 correlation ratio, 259ff., 547
 with criterion, 36
 defined, 251, 547
 between errors, 217
 between initial status and gain, 172
 mean square contingency (*c*), 238f.,
 547
 multiple, 219, 447, 547
 origin of concept, 25
 partial, 129, 219f., 547
 prediction from Spearman-Brown formula,
 223f., 566
 product moment (Pearson's r), 163, 561
 requirements for use of, 259f.
 standard error of, 267
 rank difference method, 263f., 547
 as result of overlapping elements, 289f.
 between tests and retests, 103, 163
 tetrachoric r , 266, 568
 Correlation matrix, 226, 547
 Counseling, educational, 229ff.
 vocational, 442ff.
 Creativeness, compared with craftsmanship,
 349f.
 Criminals, bodily characteristics of, 39
 intelligence of, 56f.
 Criterion, bias in, 129f.
 defined, 547
 external vs. internal, 79, 129
 multiple, 225
 reliability of, 215f.
 in tests of personal-social characteristics,
 539
 in tests for young children, 168, 310f.
 Critical ratio (C.R.), 243
 Critical score, 474, 547f.
 Cross products, 226f.
 Cross-sectional method, 32, 174, 548
 Culture-epoch theory, 32, 548
 Curve, 548
 normal (*see* Normal curve)
 percentile, 183

D

Deceit, tests of, 403f.

- Deciles, 183, 187, 548
- Deficiencies, special, 287
 measurement of, 338 ff.
 relation to behavior, 343 f.
- Deficiency (mental), American interest in, 5 ff.
 classes of, 4 f.
 definitions of, 548
 distinguished from mental disease, 3 ff.
 inheritance of, 55
 possibility of improvement, 3, 6 ff., 536 ff.
 relation to delinquency, 55 ff., 489 f.
 and social work, 486 f.
 (see also Feeble-mindedness)
- Definition, children's 44, 142, 293
 operational, 97, 559
- Degrees of freedom, (see Freedom, degrees of)
- Delayed reaction, test of, 505
- Delinquency, juvenile (see Juvenile delinquents)
- Delinquency areas, 489, 548
- Design, for classification, 115 f.
 of experiments, 283 f.
- Detroit Mechanical Aptitudes Examination, 474, 476
- Dexterity tests, 478
- Diagnostic tests, 71 f., 355 f.
- Diary records of child development, 31
- Differences, individual (see Individual differences)
- Differences between measures, fiducial limits of, 278 ff.
 formula for, 242
 significance of, 239 ff.
- Difficulty, factors determining, 142 ff.
 intrinsic vs. extrinsic, 142 f.
 order of, 141 f.
 range of, effect on test scores, 146 ff.
 of test items, 125 f, 194
- Discrete measures, 99, 549
- Discriminative ability, of test as indicated by r , 269 f.
- Discriminative value method (D.V.), 201 ff., 549
- Disease, mental, 549
 distinguished from mental deficiency, 3 ff.
 early classifications of, 27 f.
 tests in, 483 f.
- Dispersion of scores (see Range, of talent)
- Doll play, as projective technique, 422 ff.
- Dominance, lateral (see Lateral dominance)
- "Draw a man" test, 314, 316
- "Draw a man" test (*Cont.*)
 in identification of feeble-minded, 425
- Drawings, children's, 32
 by feeble-minded, 523
 as projective method, 423 ff.
- Dynamograph, 368, 549
- Dynamometer, 368, 549
- E
- Education, median extent of, 443
- Educational tests, 70 f.
 in child guidance, 464
 comparison of results with intelligence, 71, 333 ff.
 criteria in educational experiments, 337
 for the elementary school, 325 ff.
 for high school and college, 328 ff.
 predictive value, 72
 reliability of, 336 f.
- Efficiency index (Babcock), 530
- Ellis Island, testing of immigrants at, 68 f.
- Emotional stability, tests of, 78
- Emotions, expression of, 23 f.
- End effect, 379
- Environmental differences, effect on mental ability, 143, 256, 536 ff.
 relation to suitability of tests, 92, 106 f.
 tests of, 465, 488
- Episode sampling (see Sampling, episode)
- Equal-appearing intervals, method of, 378 ff., 550
- Errors, distribution of, 165, 254
 of measurement, 159
 in measurements of attitudes and interests, 383 f.
 positive vs. negative, 437
 random vs. systematic, 92 f.
 recurrent, 535 f.
 of sampling, 103, 141
- Ethical judgment tests, 77
- Eugenics, 25, 550
- Evolutionary theory, 24
- Examination, conduct of, 297 ff., 463 f.
 essay type, 332
 objective type, advantages of, 330 f.
 criticisms of, 331 f.
 preparation of, 126 ff., 332 ff.
- Examiner, mental (see Mental examiner)
- Experience, effect, on behavior, 105 ff.
 on scale values, 154 f.
 on test scores, 173 f.
- Expression, of the emotions, 23 f.
 facial, as indicator of intelligence, 35
 as indicator of personality, 38
- Extroversion-introversion, 410, 419, 550
- Eye dominance, 370 f., 372, 551

Eye movements in reading, 333, 355 ff.
 "peep hole" method of studying, 356

F

F statistic, defined, 551
 formula for, 246
 sampling distribution, table of, 248
 uses of, 249, 271 f.
 Factor analysis, 85, 225 ff., 499, 538, 551
 of personality inventories, 410, 499 f.
 of Stanford 1937 Revision, 228
 uses of, 227 f., 230 f., 290
 Factors, defined, 551
 in trait structure, 123 ff., 227 ff.
 Faculty psychology (*see* Psychology, faculty)
 Fallacies, concerning tests, 91 ff.
 naming, 102, 104 f., 139, 557
 Family resemblance, 174
 Famous men, childhood characteristics of, 512 f.
 early studies of, 13
 handwriting of, 437
 Feeble-minded persons, designation as "intellectually inadequate," 519 ff.
 distribution of, 519 ff.
 early care of, 8
 identification and classification, 3 ff., 10, 56 ff.
 IQ's, relative stability of, 170
 motor abilities of, 74
 schools for, 7 ff., 18
 (*see also* Mental deficiency)
 Fiducial limits, 278 f., 551
 "Floor" of test, 147
 Forecasting ability, index of, 164, 552
 Free association, 84
 uses of, 427 f.
 Freedom, in analysis of variance, 274 ff.
 degrees of, 234 ff., 552
 Freeman, William, trial of, 10 f.

G

g factor in mental organization, 287 ff., 552
 evidence from military psychology, 500
 Galvanometer, 402, 552
 Generosity, tests of, 81
 Genius, displayed in childhood, 512 f.
 inheritance of, 23 f., 54
 as interaction among traits, 350
 German laboratories, relation to mental testing, 30 f.
 work of, 29 f., 417
 Gifted children, characteristics of, 511 ff.
 later development of, 255, 349, 512 ff.

Gifted children (*Cont.*)
 sex differences, 245
 Graphology (*see* Handwriting)
 Group factors, 287 ff.
 Group tests, administration of, 306 f.
 factor analysis of, 228
 first use of, 67 f., 89
 item analysis and selection, 132 ff.
 Grouping, corrections for, 262
 Growth (mental), 146, 151 ff.
 age at completion of, 160
 age at midpoint, 152 f., 160 f.
 and attainment of puberty, 152
 curves of, 151, 204 ff.
 decline of, 320 f.
 early precocity of, 151 f.
 of feeble-minded, 205 f.
 prediction of, 174 f.
 rate of, 162 ff., 171 ff.
 Guilford-Martin Inventory, 450

H

Halo effect, 49, 552
 Hand preference (*see* Lateral dominance)
 Handwriting, as projective method, 436 f.
 sex differences in, 436
 "Haptic" type, 346 f., 552
 Harmonic mean, 181 f.
 Heinis Personal Constant (PC), 204 ff.
 compared with Coefficient of Intelligence, 211
 definition of, 206, 560
 used in Kuhlmann-Anderson Group Test, 210, 315 f.
 variability compared to that of I.Q., 207 f.
 Heredity (*see* Inheritance)
 Herring Revision of Binet tests, 315
 Holistic view, 431, 459, 553
 tests adapted to, 538
 Homocliticity, 260
 Homoscedasticity, 260
 Homosexuals, definition of, 553
 test performances of, 403
 Honesty in response to personality inventories, 78 ff.
 tests of, 81, 403
 Hypnosis, 553
 susceptibility to, 22, 402

I

I.E.R. Intelligence Test, 318
 Idiot, 553
 early use of term, 3
 intelligence of, 146
 mental growth curve of, 205 f.

- Idiot (*Cont.*)
 sensorimotor characteristics, 41
 Idiots savants, 350ff., 553
 possible explanations for, 352ff.
 Imagery, mental, 40
 Immigration, 68f.
 selective, as factor in test performance,
 527
 Incentives, effect on test scores, 142
 Incomes, variations in significance of,
 191ff.
 Index of Brightness, (IB), 211, 554
 Individual differences, 554
 books on, 45, 528
 Cattell's work on, 29ff., 40f.
 among children, 32f.
 Industrial selection, distinguished from
 vocational guidance, 76
 tests for, 75f.
 classes of, 473
 Industry, use of tests in, 471ff.
 Infants, predictive value of tests for, 103f.,
 310
 tests for, 50f., 308ff.
 Information tests, 70f., 326
 Inheritance, animal studies of, 506
 of artistic talent, 347f.
 Mendelian law, 53f.
 of mental traits, 13, 26
 Inkblots, use in testing, 83
 (*see also* Rorschach tests)
 Instructions, memorizing of, 303, 305f.
 for test administration, 301f.
 Intellectual inadequacy, 519ff.
 Intelligence, changes with age, 101
 correlates of, 254ff.
 different concepts of, 21, 48, 97ff., 106,
 123, 287ff., 521, 554
 early methods of appraising, 34ff.
 effects of training on, 536ff.
 patterns of, 61
 relation, to delinquency, 56ff., 489f.
 to mechanical ability, 475
 Intelligence quotient (IQ), changes in,
 165f., 312
 compared with Heinis P.C., 207f.
 constancy of, 91ff., 162ff., 312ff.
 introduction of, 62, 140
 meaning, factors influencing, 168f.
 above 180, 391
 popularity of, 212
 proportions among school children, 63
 reasons for constancy, 171ff.
 reliability of high vs. low IQ's, 170
 unequal variability of, 151, 249
 IQ Equivalent, derivation of, 199f., 206
 IQ Equivalent (*Cont.*)
 use in Wechsler-Bellevue test, 316
 Intelligence tests, 308ff.
 for the aged, 320f.
 for the elementary school, 314ff.
 for high school and college, 316ff.
 for infants, 308ff.
 for kindergarten children, 312ff.
 nonlanguage (*see* Nonlanguage intel-
 ligence tests)
 for preschool children, 311ff.
 for unselected adults, 319f.
 Interaction, in analysis of variance, 280f.,
 292, 554
 Interests, measurement of, 81, 127, 374ff.
 occupational, 76f., 280
 in school subjects, 377
 Internal consistency, 554
 as criterion for item validity, 79, 130f.
 Introversion-extroversion, 410, 419, 555
 Item analysis, 123ff., 132ff.
- J
- Juvenile delinquents, causation, multiple
 factor theory, 525
 early treatment of, 11ff.
 environment of, 489
 mental responsibility for acts, 11ff.
 mental status of, 56ff., 489f.
 treatment, experiments in, 493f.
- K
- K scale for identifying malingering, 409,
 555
 Kallikak family, 53ff.
 Knowledge, tests of, 70f.
 factual vs. understanding, 71
 Knox Cube Test, 202
 Kuder Preference Record, 447f., 478
 Kuhlmann Revisions of Binet tests, 59,
 312, 316
 Kuhlmann-Anderson Group Intelligence
 Test, 210, 316f.
 Kurtosis, 181, 555
- L
- Lateral dominance, 74ff., 555
 changes in as factors in reading diffi-
 culty, 354
 eye dominance, tests of, 370f., 372
 foot dominance, 371
 measures of, 368, 371f.
 Latin square, 285
 Leadership, variant ways of manifesting,
 123ff.
 Learning, factors in art talent, 348

Learning (*Cont.*)

- test as opportunity for, 155 f.
- Leptokurtic distribution, 181, 555
- Lie detector, 402
- Limits of testing, 61 f., 150 f., 301
- Literary talent, tests of, 348
- Logarithmic scales, 208
- Longitudinal method, 32, 555
- Lying, tests of, 81

M

- Mandible principle, 368, 555
- Manoptoscope (manopter or V-scope), 75, 370, 555
- Manual prehension, 363
- Marginal successes and failures, 61 f.
- Marital status, happiness, 402
 - relation to mental masculinity or femininity, 279 f.
- Masculinity-femininity, relation to marital status, 279 f.
 - scales for measuring, 402, 526
- Matching tests, 133 f.
- Materials for testing, arrangement of, 299 f.
- Maze tests, 504 f.
- Mean, arithmetic, 177 ff., 556
 - standard error of, 242
 - harmonic, 181, 556
- Meaning, bivariate, 126
 - of quantitative terminology, 157 f.
 - of questions, 127 f.
- Measures, physiological contrasted with physical, 154 f.
- Mechanical aptitude tests, 450, 474 ff.
 - use in army, 500
- Median, 180, 556
- Median mental age, 209 ff., 556
- Memory, Ebbinghaus's study of, 45
- Mental age, calculation of, 151
 - concept of, 11, 50 ff., 101, 140, 153
 - definition, 543, 556
 - derivation from standard scores, 200
 - as interpretative measure, 158 ff.
 - requirement for beginning reading, 323 f.
- Mental defectives (*see* Feeble-minded persons)
- Mental diagnosis, responsibility for, 12
 - social need for, 3 ff., 486 f.
- Mental disease (*see* Disease, mental)
- Mental examiner, bias in procedures, 155
 - characteristics of good examiner, 303
 - incompetence, 92 f., 121 f.
 - as technician, 120
 - training of, 56
- Mental growth (*see* Growth, mental)
- Mental organization, pattern of, 230

- Mental set, emphasized by Binet, 521 f.
- establishment of, 301 f., 305

Mental tests, 556

- Cattell's series for college students, 13, 41
 - early tests for children, 41 ff.
 - and scientific research, 65, 504 ff.
- Merrill-Palmer test, 169, 178, 312, 314
- Mesokurtic distribution, 181, 556
- Mind-body problem, 28
- Minnesota Occupational Classification, 66
- Minnesota Preschool Scales, 115, 298, 312 f., 316
- Mode, 181, 556 f.
- Molar view of mental processes, 557
 - opposed to molecular view, 20, 48
 - Rorschach test as example, 84
- Morality tests, 77
- Motivation, effect on test performance, 104
 - in tests of personal-social characteristics, 391 f.
- Motor abilities, correlations with Binet tests, 364
 - correlation with other abilities, 371 f.
 - effect of practice, 360 f.
 - intercorrelation of, 360 f.
 - tests of, 73 f., 362 ff.
- Multiphasic Inventory (Minnesota), 408 ff., 557
- Multiple-choice tests, 71, 98, 133 ff., 557
- Muscle tensions, cues for response, 155
- Musical ability, tests of, 344 ff.
- Musical education, measurement of progress, 345

N

- Naming fallacy (*see* Fallacies, naming)
- National Intelligence Test, 256
- Negative phase, 311
- Negroes, 528
- Nervous habits, 395 f.
- Neural growth, precocity of, 160
- Nondirective interviewing, 460, 558
- Nonlanguage intelligence tests, correlation with Stanford-Binet, 69
 - development of, 68 ff.
 - effect of practice on, 155 f.
- Minnesota Preschool, 312
 - of personality and conduct, 81
- Pintner's, 316
 - predictive value of, 312
 - use with immigrants, 68 f.
 - with adolescents and adults, 69
 - validity of, 67
- Wechsler-Bellevue, 318
- Nonsense syllables, 45, 558

- Normal curve, 25, 147, 189, 558
 conformity of test scores to, 149, 200
 transmuting data to form of, 190ff.
- Normalizing, 149, 190ff., 558
- Norms, 109, 558
- Null hypothesis, 136, 232ff., 278, 558
- O
- Objective examinations (*see* Examination, objective type)
- Occupational interests, of gifted subjects, 515
 tests of, 76f., 280, 447f., 478
- Occupations, classification, 66
 lists of, 444
 paternal, use in test standardization, 66, 114
- Ohio State University Psychological Test, 318
- "Only" children, language development of, 143
- "Optic type" of graphic expression, 346
- Orthogonal traits, 290
- Oseretzky tests, 74, 364f.
- Overlap, basis of correlation, 289f.
 in Kuhlmann-Anderson Group Test, 316
 between successive testings, 171f.
 between tested groups, 241
- Overstatement tests, 403ff., 559
- P
- Paper Form Board Test (Minnesota), 450, 474, 476
- Parent-child relationships, through doll play, 422f.
 measurement of, 405f., 464ff.
- Part-whole relationship, 172ff., 327
- Paternal occupation (*see* Occupation, paternal)
- Per cent placement, method of, 208f., 559
- Percentile rank, comparison with IQ, 184f.
 comparison with per cent placement, 208
 definition, 182f., 559f.
 limitations and hazards, 184ff.
 relation to standard scores, 189
- Perceptual abilities, tests for, 73f.
- Performance tests, of educational achievement, 325
 of intelligence (*see* Nonlanguage intelligence tests)
 for prediction of conduct, 403f.
- Permillés, 188, 560
- Personal attractiveness, effect of, on estimates of ability, 34
- Personal constant (P.C.) (*see* Heinis Personal Constant)
- "Personal equation," 30, 560
- Personality, Cattell's analysis of, 411ff.
 definitions, 390, 560
 of industrial workers, 479f.
 inheritance of, 24
 significance for vocational guidance, 450f.
- Personality inventories, 406ff., 560
 criticisms of, 78f., 406f., 480
 diagnostic value of, 81
 origin of, 495
 types of, 77f.
 use in Army Air Force, 498ff.
 wording of, 127f., 407ff.
- Phi coefficient, 499, 501, 560
- Philosophy of education, 326f.
- Phrenology, belief in, 68
 contribution to mental testing, 40
 definition of, 560
 history of, 38f.
- Physical growth, types of, 159
- Physically handicapped, children, 468 f., 529 f.
 war veterans, 467
- Physiognomy, 34ff., 561
 judgment of character by, 438
- Physique, and intellect, 36ff., 257f.
- Pintner-Paterson Performance Scale, 210
- Pitch of tones, individual differences in hearing, 26f.
 of voices in testing, 301, 305
- Platykurtic distribution, 180f., 561
- Plythsmograph, 402, 561
- Point scale, defined, 561
 distinguished from year scale, 59, 153
 Yerkes Point Scale, 59ff.
- Position, of examiner and subject, 299, 304
- Practice, effect of, 154f., 159, 360
- Praise, effect on test scores, 142
 variations in amount of, 155, 302, 305
- Prediction, from infant tests, 103f., 310
 of later from earlier test standing, 168, 173; 312
 by regression equation, 206
- Preferences, by Kuder Preference Record, 447f., 478
 measurement of, 381ff.
- Prehension, manual (*see* Manual prehension)
- Primary abilities, 85, 112, 229f., 314, 533, 561
- "Private worlds," 81
- Probable error (P.E.), 201, 561
- Probability, of changes in IQ, 91 ff.
 of changes in sign of correlation coefficient, 136

- Probability (*Cont.*)
 expression of, 233
 inferred from correlation, 252f., 270
 integral, 168
 statistics of, 85, 166f.
- Probability curve (*see* Normal curve)
- Probability integral, table of, 195
- Product moment method (*see* Correlation)
- Product scales, 70, 325f., 348, 561
- Profile, psychological, 100, 339ff., 561
- Projective methods, 81ff., 415ff., 534, 562
 critical discussion of, 440f.
 essential feature of, 83
 types of, 82
- Proof, burden of, 537
- Psychogram (*see* Profile, psychological)
- Psychologists, 562
 in clinics, special problems of, 469f.
 in industry, 481f.
 in mental hospitals, qualifications of, 485f.
 number engaged in testing, 90
 training of, 91
- Psychology, abnormal, 21f., 23, 543
 associationism, 21, 544f.
 "depth," 84
 faculty, 20, 43, 551
 and medicine, 483f.
- Puberty, relation to test performance, 152
- Public opinion poll, 81, 120, 562
 area method, 387f.
 effect of refusals, 386
 factors inducing bias in, 118f.
 method of, 385f.
- Purification, of test universe, 112, 229f.
- Q
- "Qualitative," 562
 distinguished from "quantitative," 147
 qualitative aspects of behavior, 293
 qualitative differences in test performance, 320f.
- Quartiles, defined, 188, 562
 quartile difference, as criterion for item selection, 130, 135f.
- Questionnaire, 563
 adjustment, 78, 406f.
 early use of, 31f., 46
 sent by mail, returns from, 388
 in studies of lateral dominance, 370f.
 wording of, 127f., 133f.
- Quotient method, 175f.
- R
- Racial characteristics, criteria for estimating personality, 38
- Racial characteristics (*Cont.*)
 intellectual, 527f.
 use of TAT in determination of, 431
- "Random," defined, 563
- Random errors, 92f.
- Random numbers, 114
- Random sampling (*see* Sampling, random)
- Range, 563
 of difficulty, 125f., 141, 147ff.
 of talent, effect on correlation, 68, 138, 163f.
 of test content, 124f.
- Reaction time, 563
 as measure of intelligence, 40
 as measure of manual speed, 367
- "Readiness" tests, 72f., 323ff., 563
- Reading difficulties, 72, 332f., 354f.
 diagnostic tests for, 355f.
- Reading tests, factors influencing performance, 356
 types of, 71
- Recapitulation, theory of, 32, 563
- Recidivists, 490f., 563
- Recording, methods of, 416f.
- Regression, 563
 Burt's equation, 253
 curvilinear, 260f., 289
 effect on accomplishment quotient, 335
 equation, 206
 multiple, 220f.
 rectilinear, 259f., 289
- Reification, definition, 107, 564
- Reliability, compared to validity of test, 46, 79, 106
 definition of, 106, 213, 564
 emphasis on, 68
 of high vs. low IQ's, 170
 significance of, 138
 three ways of determining, 213f.
- Representative sampling (*see* Sampling, representative)
- Reproof, effect on test scores, 142
- Residuals, in factor analysis, 227f.
- Resistance, effect on test score, 166
- Retardation, school (*see* School retardation)
- Rigidity, mental, of feeble-minded, 522f.
- Rorschach test, 432ff., 564
 criticism of, 84f.
 deals with signs, 97f.
 among feeble-minded, 523
 group form, 436
 history of, 432
 movement responses, 485
 scoring, 83ff.

- Rorschach test (*Cont.*)
 use in army, 500f.
- Rosenzweig Picture Frustration Test, 438f.
- S
- s factors in mental organization, 287f., 564
- Samples, distinguished from signs, 82, 102⁹
 as predictors of behavior, 100f.
 random, in analysis of variance, 272ff.
 representative vs. random, 101f.
 required proportions, 113
 small, statistical methods for, 85, 243ff.
 source of reference, 117f.
 test items as samples, 126ff., 167
- Sampling, area sampling, 387ff.
 basis of most tests, 86, 106
 in Binet tests, 42ff.
 episode sampling, 392ff., 394ff., 550
 errors of, 103, 141
 problems of, 109ff.
 purpose of, 109
 random, 113f., 116
 stratified or representative, 114f.
 of subjects for Stanford-Binet tests, 64ff.
 and test interpretation, 120f.
 in tests of personality and conduct, 81
 time sampling, 392ff., 568
 limitations of, 397f.
- Scale values, 70
 determination of, 130f., 141ff.
 effect of differing experience on, 154f.
- Scatter diagram, 259, 565
- Schizophrenia, confusion with mental deficiency, 5
 drawings in, 424
 Rorschach tests of, 485
 mental deterioration in, 530
- School retardation, early statistics, 14ff., 52f., 322f.
 effect on child behavior, 16f.
 among immigrants, 68f.
 regional differences, 17f.
- Score, meaning of, 98ff.
- Scoring keys, multiple, 76, 78
- Scottish children, intelligence of, 245
- Seguin Form Board, 202
- Semantics, 127f., 565
- Sensation, psychology of, 29ff., 565
- Service and self control, tests of, 403
- Sex differences, 526ff.
 in interests, 76f.
 in mental traits, 13
 in motor abilities, 365ff.
 in test performance, 245, 249
 in variability, 245
- Shirley's twenty-five babies, 362
- Shirley's twenty-five babies (*Cont.*)
 later development of, 364
- Significance, level of, defined, 136, 233, 565
- Signs, bodily, of mental traits, 37ff.
 in classification of mentally defective, 523
 defined, 565
 derived from Wechsler-Bellevue tests, 318, 447, 523
 distinguished from samples, 82, 416
 as predictors of behavior, 100
 in Rorschach test, 432f.
 in word association tests, 428f.
- Simple structure, 228, 290
- Size, physical, effect on judgments of ability, 34
- Skewness, 566
 effect on standard error of percentile, 187
 resulting from insufficient range of difficulty, 147ff.
 transforming to normal distribution, 190ff.
- Social agencies, use of mental tests by, 486f.
- Social participation, 390ff.
 methods of study, 391ff.
- Sociogram, 399f., 566
- Sociometry, 398ff., 566
- Spacing, of test items, 141ff., 145f.
- Spearman-Brown prophecy formula, 223ff., 566
 requirements for use of, 224
 in time-sampling studies, 394
- Special classes, New Jersey law, 56
 purposes of, 19
 for retarded children, 18, 52ff.
- Speech, as means of studying social behavior, 396
 neural mechanism in, 369
 relation to intellect, 4
 of twins vs. "only" children, 143
- Speed, rate of, in test administration, 301
 relation to accuracy, 71, 327
 relation to intelligence, 40, 69f.
 test of, 181f.
- Standard deviation (S.D.), defined, 163, 566
 distinguished from standard error, 180
 of IQ distribution, 168f.
 standard error of, 246ff.
- Standard error (σ), defined, 179f.
 of mean, 271
 of percentage, 187
 of percentile, 187
 of true score, 166, 214f.

Standard scores, 149
 applications, 194 f., 336, 340 ff.
 defined, 189, 566
 growth in popularity, 212 f.
 Standardization group, changes in, 154 f.
 defined, 566
 universe defined by, 106, 109
 Standards, reference, 140 ff., 338 f.
 uniform, as basis for new tests, 153
 U.S. Bureau, 153 f., 155
 Stanford 1916 Revision, defects of, 66
 intelligence quotients, constancy of, 91 ff.
 standardization of, 62 ff., 140, 162
 Stanford 1937 Revision, ages for which
 best suited, 315 f.
 factor analysis of, 228
 standardization of, 65 f., 115
 variability of IQ's, 169, 199
 Statistical methods, application to testing,
 68, 85 f., 108
 for defining a universe, 111
 first application to biological data, 24 ff.
 Stimulus, nature of, 418
 Stratified sampling (*see* Sampling, stratified)
 Strength, muscular, tests for, 367 f.
 Strong Vocational Interest Test, 280, 447 f.,
 478
 performance of gifted children on, 515
 Stuttering, theories regarding, 369 ff.
 Superior children, difficulties of adjustment, 391
 Terman's interest in, 65
 (*see also* Gifted children)
 Sympathetic behavior, 394 f.
 Sympathetic magic, 139, 567
 in projective method, 425

T

t statistic, 567
 formula, 243 f.
 in formula for determining significance
 of a correlation, 268
 table of, 244
 T-score method, 196 ff., 567
 Tachistoscope, 355, 567 f.
 Talents, special, 287
 measurement of, 338 ff., 344 f.
 motor, 362 f.
 Tapping, rate of, as measure of hand
 dominance, 370
 Taussig Industrial Classification, 115 f.
 Tautophone, 437
 Teachers, training of, New Jersey law,
 56
 in use of tests, 456

Teachers (*Cont.*)
 at Vineland, 53 ff.
 Teachers' judgments, of pupils' ability,
 35 f., 41 ff., 191, 221
 Tension systems, 292 f.
 Terman Group Test, 318
 Terman-McNemar Test of Mental Ability,
 318, 447
 Testing, and scientific investigation,
 504 ff.
 Testing movement, growth of, 89 ff., 121
 Testing program, aims of, 457
 for high school students, 329 f.
 Testing room, arrangement of, 298 f.,
 305
 Tetrad differences, 226 f., 286 ff., 568
 Thematic Apperception Test (TAT),
 428 f., 568
 group forms, 430
 use in army, 500
 Thinking, abstract, 21, 106
 deficiency in, 353 f.
 and organic brain damage, 531
 Thinking aloud, 79 ff., 258
 Three hole test, 370
 Tonometer, 345, 568
 Tonoscope, 345, 568
 Topology, 292 f., 538 f.
 Trait, definition of, 100 f., 568
 names, 104 f., 132, 292, 411
 relation to sampling theory, 109 ff.
 source vs. surface, 413, 566
 True score, definition and formulas, 166,
 568
 estimated, 340
 True-false tests, 133
 Twins, resemblance of, 24
 speech of, 143
 Two-factor theory, 226 ff., 286 ff.

U

Understanding, measurement of, 326 f.,
 332
 Units of measurement, 140 ff.
 equal, 144
 psychological vs. physical, 45, 154 f.
 spacing of, 141 f.
 Universe, 569
 aspects of, 110 ff.
 behavioral, 99 ff., 569
 boundaries of, 111 f., 119 f., 190
 derivation of word, 111
 means of defining, 97 ff.
 need for defining, 119, 131 f., 134
 test as sample, 106 f.
 Unmarried mothers, study of, 488 f.

V

- V-scope (*see* Manoptoscope)
 Valence, 293
 Validity, definition of, 106, 213, 569
 determination of, 36, 101
 facts needed for determining, 439f.
 relation to reliability, 46, 68, 106, 222f.
 of tests for industrial workers, 480f.
 Values, in personal-social behavior, 391
 tests of, 375 ff.
 Variability, 569
 basis for establishing zero point, 146
 effect of change in, 169
 effect on correlation, 68, 138, 163 ff.
 of IQ distribution, 163f.
 irregularities in, 151
 significance of, 25
 Variables, dependent vs. independent,
 253 ff.
 x and y , 259
 Variance, analysis of (*see* Analysis of variance)
 definition of, 271, 569
 between and within groups, 271 ff.
 Variance ratio (*see* F statistic)
 Verbal summator (*see* Tautophone)
 Vineland Social Maturity Scale, 405
 Vineland Training School, 51 ff.
 summer school, 53
 Visual apprehension, span of, 45, 569
 Vocabulary, records, 31
 relation to mental age, 4
 Vocational counselor, qualifications of,
 444 ff.
 Vocational guidance, 442 f.
 distinguished from industrial selection,
 76, 442
 tests for, 75 f.
 classification of, 445 f.
 vocational interest tests, 280, 447 f.
 Voice, pitch of, in testing, 301, 305
 quality, measurement of, 345 f.

W

- w factor in mental organizations, 287
 among gifted children, 517 f.
 Wechsler-Bellevue Test, 316 ff.
 predictive value, 463
 "signs" derived from, 403, 447
 Weighting, equal, 137
 Guttman's method, 130 f.
 as means of normalizing a distribution,
 149
 multiple correlation as basis for, 219 f.
 negative vs. positive, 136 f.
 in TAT, 431 f.
 Wild Boy of Aveyron, 6 ff.
 Wishful thinking, 39
 Word lists, Thorndike's, 144
 Work decrement, measurement of, 182
 World Test, 439
 World War I, development of trades tests,
 75
 and group testing, 67 f., 89, 494 f.
 mental level of soldiers in, 160
 World War II, psychological testing in,
 495 ff.

Y

- Year scale, distinguished from point scale,
 59
 inequalities in, 178 f.
 placement of items in, 150 f.
 requirements of, 74

Z

- z function (Fisher), 569
 compared with Pearson's r , 267 f.
 standard error of, 270
 table of, 268
 Zero point, in Discriminative-value
 method, 201
 location of, in tests, 86, 145 f., 153
 in T-score method, 196 f.
 Zest, as factor in accomplishment, 514 ff.